PolyNC: a natural and chemical language model for unified polymer properties prediction

Haoke Qiu, [ab] Lunyang Liu, [**a] Xuepeng Qiu,[bc] Xuemin Dai,[c] Xiangling Ji, [ab] Zhao-Yan Sun [*ab]

[a] State Key Laboratory of Polymer Physics and Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China. E-mail: lyliu@ciac.ac.cn (L.Y.L.), zysun@ciac.ac.cn (Z.Y.S.)

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

[c] CAS Key Laboratory of High-Performance Syntheic Rubber and its Composite Materials, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China.

Supplementary Information

Table of contents

S1. Distribution of Token Sizes

The input instructions (natural language and chemical language inputs) are tokenized at the character level, dividing them into natural language tokens and chemical tokens. This tokenization strategy has been proven to provide better performance and stronger expressive capabilities. The distributions of token sizes in the training and validation sets are depicted in **Figure S1**, from which we determined to set the input token size to 150 and the output token size to 8 to strike a balance between memory requirements and computational speed.

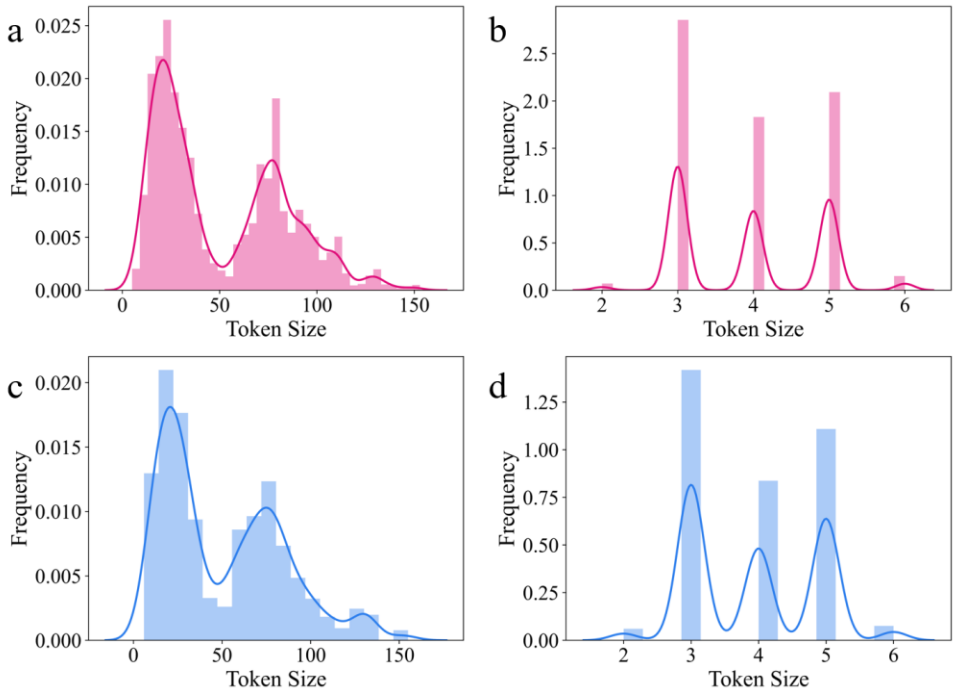


Figure S1. Distributions of token sizes. Subplots **a**, **b**, **c**, and **d** respectively represent the distribution of token sizes for the training set input, training set output, test set input, and test set output.

S2. Fitting Curve of PolyNC

PolyNC exhibits outstanding performance in regression tasks, as illustrated in **Figure S2**. Despite significant differences in the ranges of values for each property, remarkably, PolyNC achieves the best performance in tasks where the properties are not explicitly differentiated. This outcome provides evidence of PolyNC's potential understanding of diverse chemical knowledge across multiple properties.

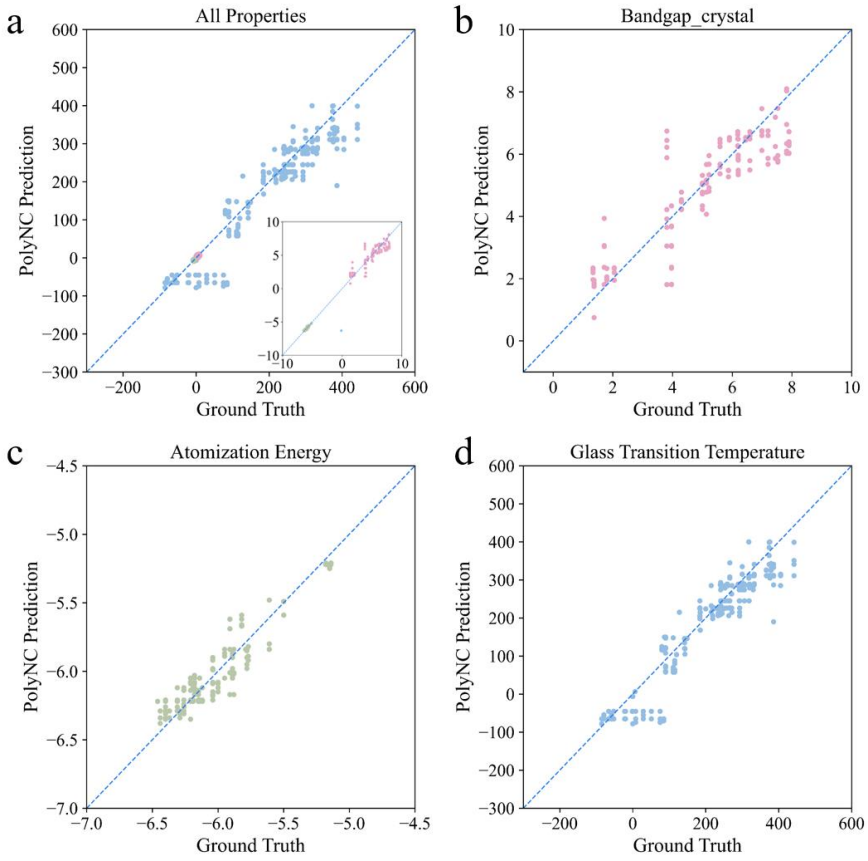


Figure S2. Fitting plots of PolyNC on each regression task.

S3. Performance of ML baselines

S3.1 Regression tasks

The evaluation metrics used for the regression models were R^2, MAE, and MSE, as shown in Figure 3 of the main text. The values for MAE and MSE can be found in the table below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

where n is the number of samples, y_i represents the ground truth values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ represents the average value of the ground truths.

We divided the dataset with training set/test set = 0.8/0.2, and then all models were trained and evaluated on the same training and test sets. Since the data in the training and test sets were divided homogeneously (Figure 3a in the main text), no cross-validation was performed to reduce the overhead of computational resources and time. The detailed results of MAE and MSE are as follows.

Table S1. MAE and MSE of each regression model.

Model	T_g		BC		AE		All	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	111.20	17299.20	0.99	1.84	0.28	1.77	107.18	16044.68
SVR	106.07	18433.34	1.70	4.89	0.27	0.11	84.34	21151.63
GPR	188.03	50253.78	5.14	30.83	5.96	35.79	81.37	20308.13
GCN	44.62	3146.62	0.55	0.51	0.37	0.15	26.79	1763.73
RF	40.48	3671.94	0.80	1.00	0.05	0.01	27.31	2591.18
BAG	30.64	2837.74	0.65	0.77	0.04	0.00	25.77	2438.35
RR	32.17	1884.57	0.65	0.78	0.05	0.01	37.01	2632.58
ADA	35.93	2276.52	0.67	0.75	0.09	0.01	34.87	2347.12
EXT	30.59	2633.18	0.63	0.63	0.04	0.00	22.85	1561.41
PolyT5	37.68	2552.54	0.59	0.67	0.09	0.01	15.14	1013.10

S3.2 The HRC Classification task

The evaluation metrics used for the classification models were Accuracy, Precision, Recall, and F1 Score, as shown in Figure 4 of the main text. The values for Accuracy, Precision, Recall, and F1 Score can be found in the table below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

$$\text{weight}_i = \frac{C_i}{\sum_{i=1}^{3} C_i} \quad (8)$$

where TP represents true positive, TN represents true negative, TP represents false positive, FN represents false negative, and C_i represents the number of class *i*.

The performance of eight baseline models for classification tasks, including Logistic regression (LRC), Naive bayes (NBC), Support vector machine (SVC), AdaBoost (ADAC), Decision tree (DTC), Random forest (RFC), K-nearest neighbors (KNNC) and XGBoost (XGBC), is depicted in **Table S2** and **Figure S3**. In this classification task, LRC and SVC performed relatively poorly, resulting in cross-class errors. The remaining ML models performed slightly better, without generating cross-class errors, but their performance in specific categories was weaker than PolyNC.

Table S2. Accuracy, Precision, Recall and F1 Score of the classification models.

Model	Accuracy	Precision	Recall	F1 Score
LRC	0.08	0.01	0.08	0.01
NBC	0.35	0.12	0.35	0.18
SVC	0.57	0.32	0.57	0.41
ADAC	0.49	0.79	0.49	0.43
DTC	0.68	0.64	0.68	0.65
RFC	0.73	0.67	0.73	0.70
KNNC	0.70	0.73	0.70	0.70
XGBC	0.73	0.67	0.73	0.70
PolyT5	0.82	0.82	0.82	0.81

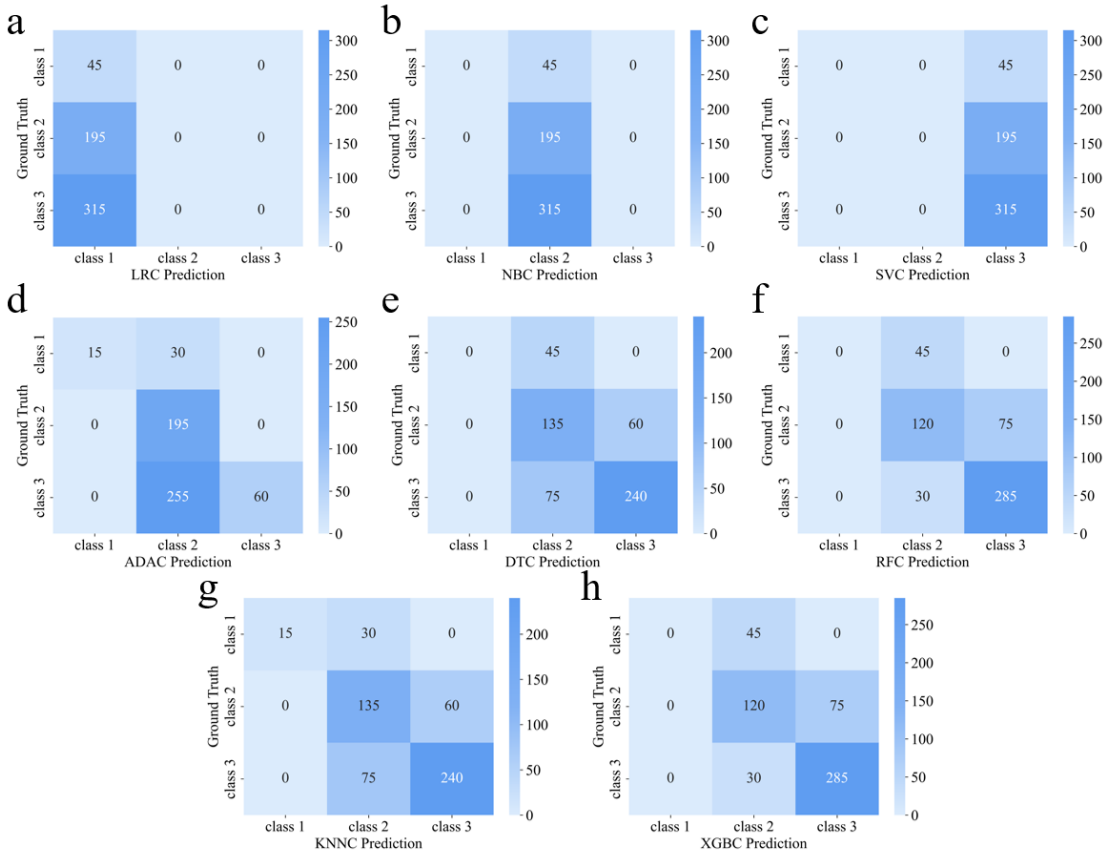


Figure S3. The performance of eight baseline models for classification tasks. (a) Logistic regression (LRC). (b) Naive bayes (NBC). (c) Support vector machine (SVC). (d) AdaBoost (ADAC). (e) Decision tree (DTC). (f) Random forest (RFC). (g) K-nearest neighbors (KNNC). (h) XGBoost (XGBC).

S4. Synthesis and Characterization

2,3,3',4'-biphenyl tetracarboxylic diandhydride (3,4'-BPDA) were purchased from Shanghai Chemical Reagent Plant and used directly. 4,4'-Diamino-p-terphenyl (DPTP) and 4,4'-benzidine (Bz) were purchased from Changzhou Sunlight Medical Raw Material Co. Ltd. and purified by sublimation before use. Precursors of PI, polyamide acid (PAA), were synthesized from the reaction of 3,4'-BPDA with aromatic diamines (DPTP, Bz) in DMAc, respectively.

Fourier transform infrared (FTIR) spectra were acquired using a VERTEX 70 spectrometer, covering the range of 4000 to 400 cm-1. The FTIR analysis revealed distinct peaks at approximately 1776 cm-1 and 1708 cm-1, corresponding to the asymmetric and symmetric stretching vibrations of the C=O bond in the imine ring,

respectively. Additionally, a peak at 1369 cm-1 indicated the stretching vibration of the C-N bond in the imide ring, while a characteristic peak at 1500 cm-1 represented the benzene ring. The infrared spectra analysis (**Figure S4**) of the copolymerized polyimide samples confirmed the presence of imine carbonyl absorption peaks, which are indicative of the imide ring structure. These findings provide compelling evidence for the successful synthesis of polyimide across all sample proportions.

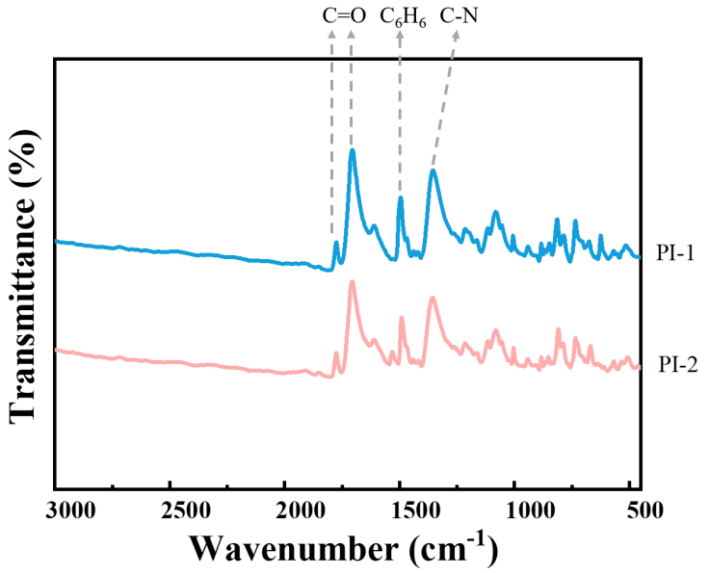


Figure S4. FTIR spectra of PI-1 and PI-2.

To further validate specific functional groups and ensure the successful synthesis of the polyimides, we employed triple quadrupole gas chromatography/mass spectrometry (GC/MS) analysis (**Figure S5**, **S6**). With FTIR and GC/MS analysis, we found that all structures were successfully synthesized.

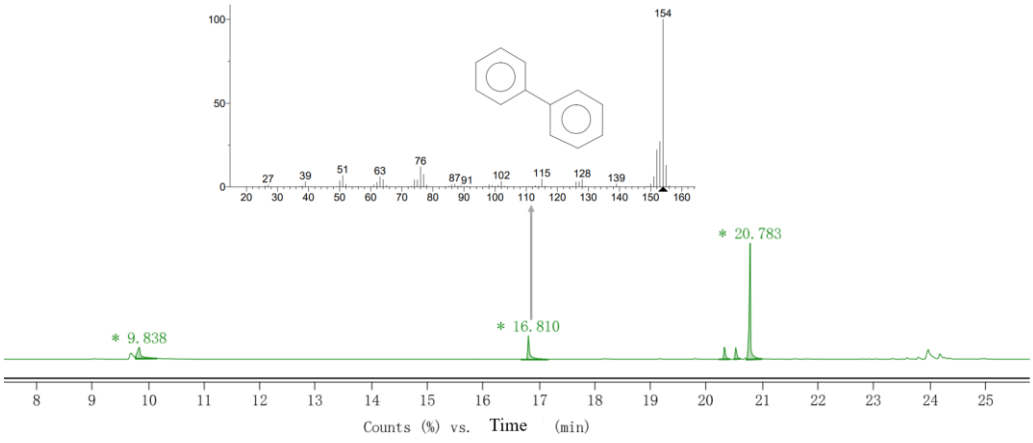


Figure S5. GC/MS of PI-1. The characteristic functional group of PI-1, diphenyl, can be identified.

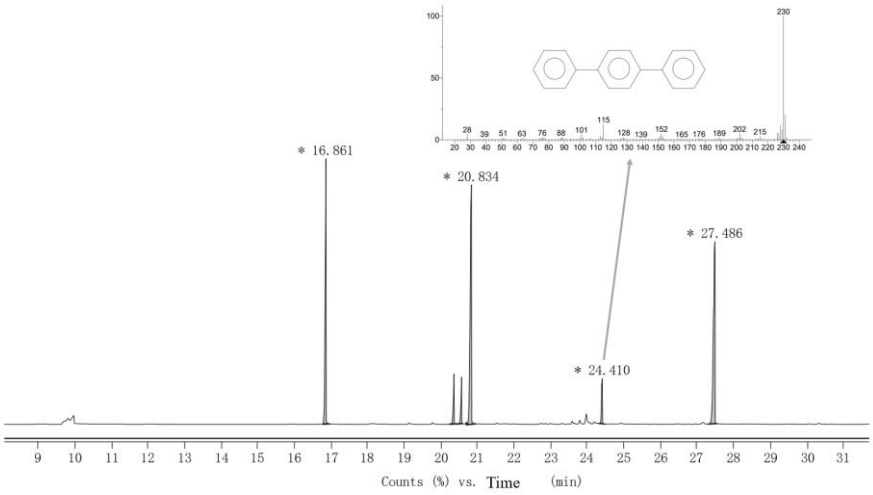


Figure S6. GC/MS of PI-2. The characteristic functional group of PI-2, triphenyl, can be identified.

The inherent viscosities (η_{inh}) of PAA solutions were measured with an Ubbelohde viscometer at a concentration of 0.5 g/dL in DMAc at 30 °C. DSC analysis was performed using a calorimeter (Q20 DSC, TA instruments), and the sample mass was approximately 3 mg and thus obtained the T_g of each PI. A summary of Inherent viscosity and T_g is listed in **Table S3**. PolyNC predicted the least deviation (5°C and 20°C) between predicted and ground truth values for these two samples compared to other baseline models of the main text.

Table S3. Inherent viscosity of the PAA solution of PI-1 and PI-2.

Sample	η_{inh} (dL/g)	T_g (ºC)	PolyNC's Prediction (ºC)
PI-1	1.82	405	410
PI-2	1.65	390	410

S5. Attention Analysis

The encoder of PolyNC consists of 12 attention heads, as shown in **Figure S7** and **S8**, each focusing on different contexts to extract distinct knowledge. The attention scores for the fifth and ninth attention heads of PI-1 and PI-2 are shown and detailed in Figure 5 of the main text. It can be observed that the attention head 1, 2, 9 and 11 mainly focuses on the tokens themselves. The attention head 3, 4, 5, 8, 10 primarily attends to adjacent tokens for each token because the nearby tokens in polymer SMILES usually represent atoms bonded to each other in the polymer, and atoms are most significantly affected by their local environments, while the attention heads 6, 7 and 12 focus on more global information. The simultaneous perception of self-information, neighboring information, and global information contributed to the success of PolyNC. The attention scores of PI-1 and PI-2 exhibits similar attention scores matrices due to the structural similarity between them, which implies that learning more data has the potential to enhance PolyNC's discriminative capabilities.

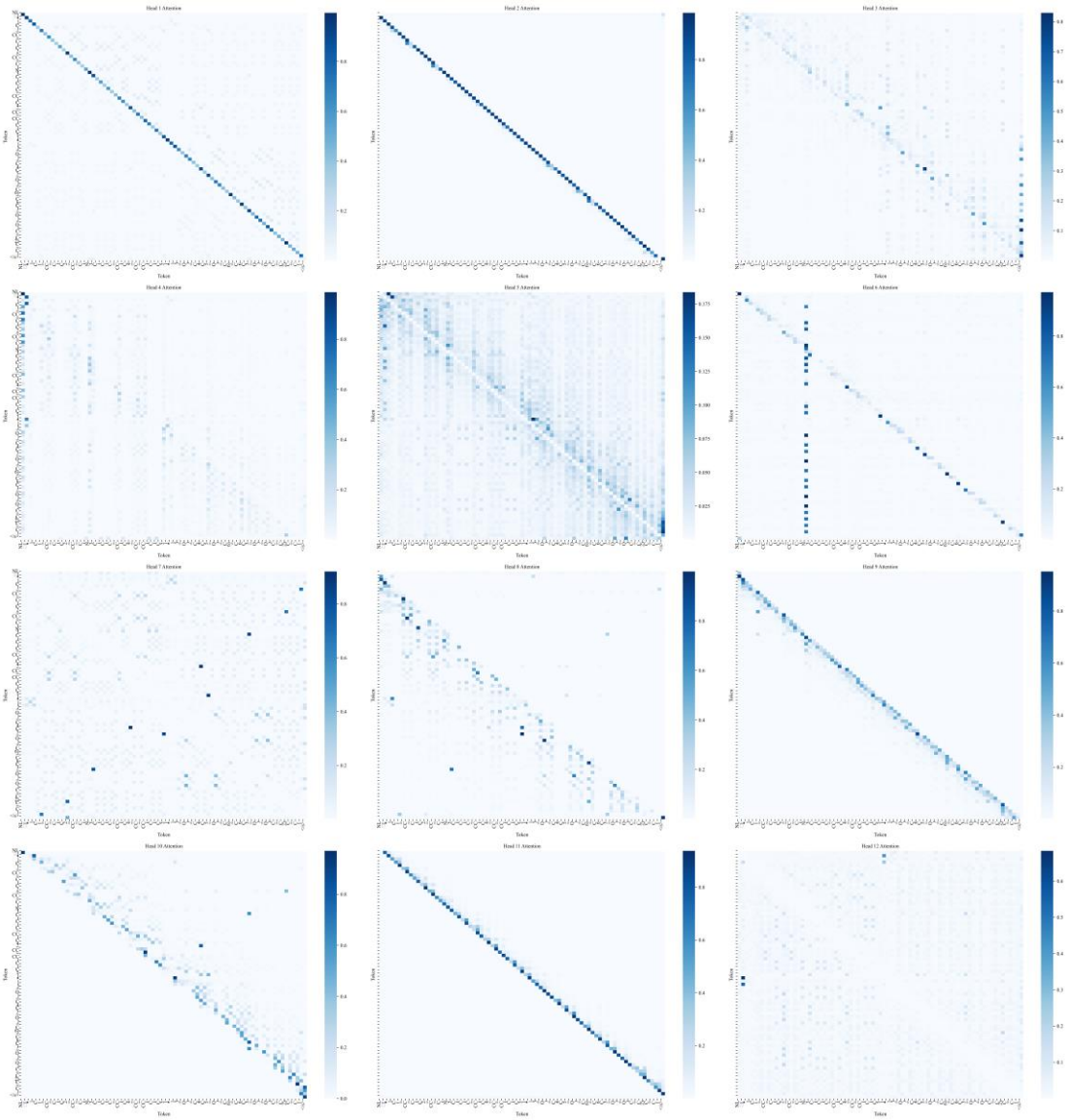


Figure S7. Attention maps of PI-1. Prompt is 'Predict the Tg of the following SMILES: [*]C1=CC=C(C2=CC=C(N3C(C(C=CC(C4=CC=CC(C(C5[*])=O)=C4C5=O)=C6)=C6C3=O)=O)C=C2)C=C1'

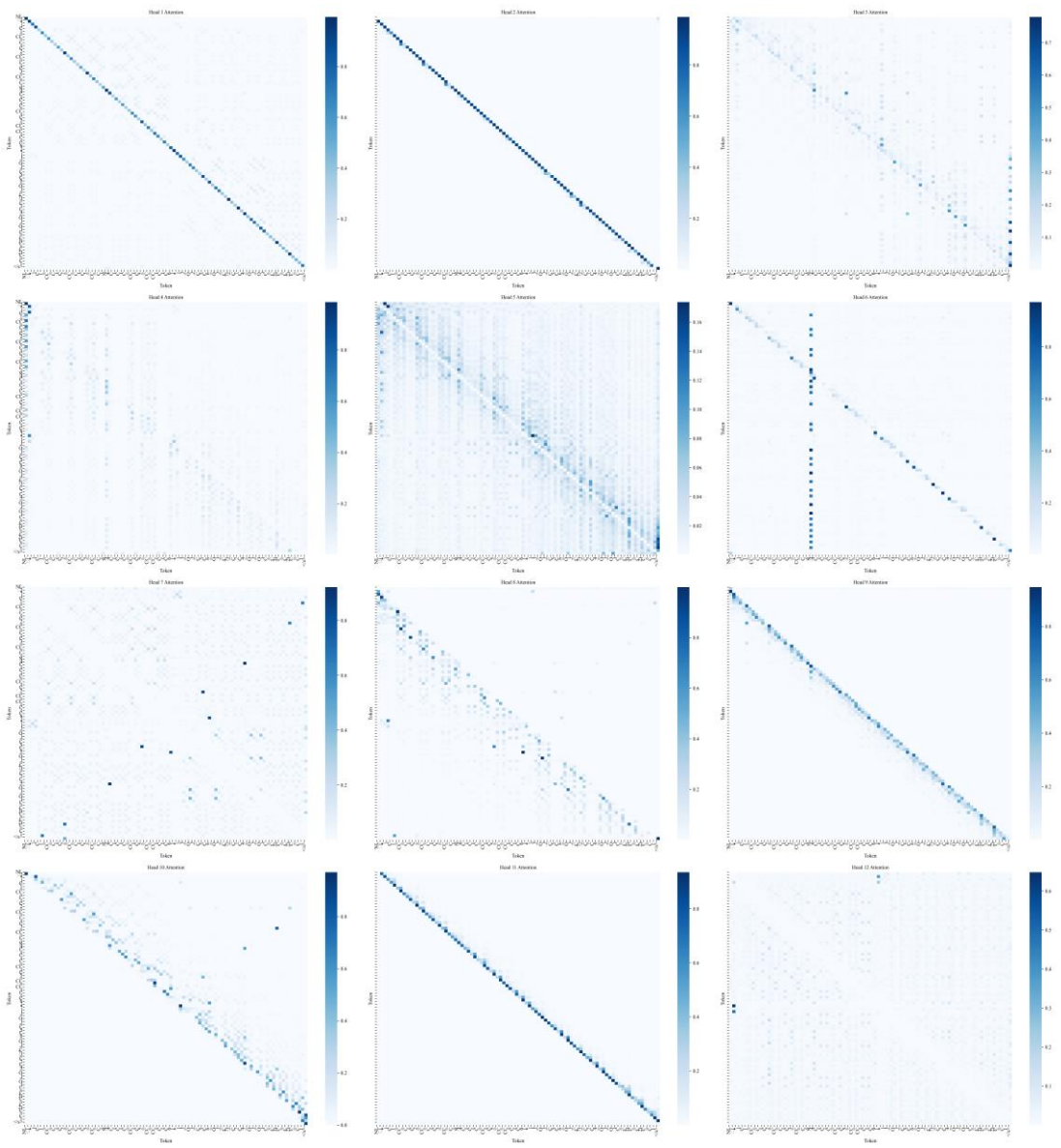


Figure S8. Attention maps of PI-2. Prompt is 'Predict the Tg of the following SMILES: [*]C1=CC=C(C2=CC=C(C3=CC=C(N4C(C(C=CC(C5=CC=CC(C(C6[*])=O)=C5C6=O)=C7)=C7C4=O)=O)C=C3)C=C2)C=C1'

S6. Comparison with Current Large Language Models (LLMs) and LLMs-induced Models

S6.1 Comparison with ChatGPT, Claude-instant, Llama-2-70b, and Google-PaLM

To demonstrate the irreplaceability of PolyNC in accelerating polymer development, we compared it with the most advanced general-purpose large models available today, including ChatGPT, Claude-instant, Llama-2-70b, and Google-PaLM. The results are shown in **Figure S9**. In the realm of polymer research tasks that pique our interest, the performance of PolyNC surpasses that of extensively trained large predictive models, which require substantial human resources, computing resources, and financial investments, highlight the superiority of our model. Based on the same prompts, our model provides more precise and professional answers. Although PolyNC may provide fewer natural language responses compared to larger commercial models, its answers are more direct and accurate. Furthermore, any shortcomings in its responses can be easily addressed by refining the training corpus, which does not diminish the success of PolyNC in predicting a wide range of polymer properties based on both natural and chemical language. In addition, PolyNC requires significantly less memory space and hardware resources compared to the aforementioned larger commercial models. This highlights the advantages of domain-specific models in terms of resource efficiency.

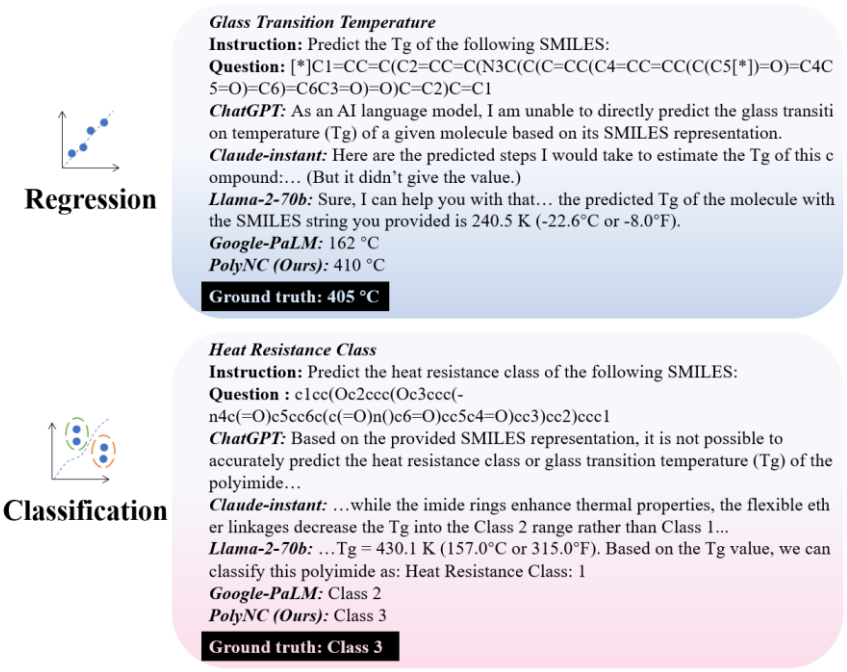


Figure S9. Each LLM's response to the polymer-related prompt. Polymer-specific PolyNC provides more precise and professional answers. Due to the custom division of heat resistance levels in polyimides based on previous experience, we not only provide the depicted prompt to the general large-scale models but also inform them about the criteria for the divisions for the elimination of unnecessary errors with "The heat resistance level is distinguished based on Tg, where: Class 1 is for polyimides with a Tg greater than 400°C. Class 2 is for polyimides with a Tg between 300-400°C. Class 3 is for polyimides with a Tg less than 300°C".

S6.2 Comparison with TransPolymer and polyBERT

In the manuscript, we have demonstrated that PolyNC can achieve superior or comparable performance compared to descriptor-based or graph-based ML models. Additionally, we have also compared the performance of PolyNC with two state-of-the-art LLMs-induced models (TransPolymer, npj Comput Mater 9, 64 (2023) and polyBERT, Nat Commun 14, 4099 (2023)) that have been reported recently.

In the first place, both TransPolymer and polyBERT employ a paradigm based on pre-training and fine-tuning to predict polymer properties. By means of pre-training, a feature extractor capable of extracting polymer descriptors is obtained, and the extracted machine fingerprints are subsequently utilized as inputs for downstream neural networks. However, PolyNC takes a one-step, end-to-end and multi-tasks approach for achieving the same objective, which is one of the advantages of PolyNC.

The key aspect in predicting polymer properties using language models lies in extracting appropriate latent representations for polymers. This step typically requires a large amount of data (5M for TransPolymer and 100M for polyBERT). In contrast, PolyNC successfully extracted these patterns using supervised learning with just over 20,000 data samples (**Figure S10**).

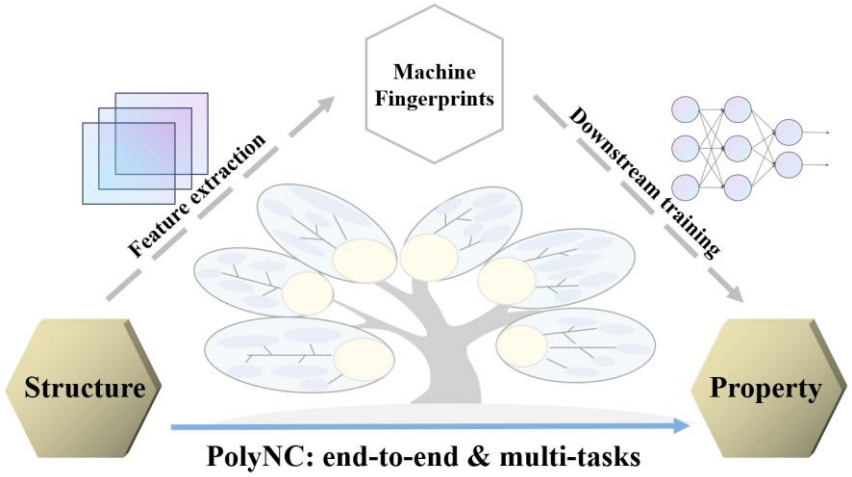


Figure S10. PolyNC can achieve end-to-end multitask prediction based on both natural language prompts and chemical language prompts.

While expanding the functionality of LLMs in polymers domain is the focus of our research, we have also compared the performance of PolyNC with TransPolymer and polyBERT, which are also derived from LLMs. According to their papers, TransPolymer and polyBERT demonstrated comparable model performance. Due to the availability of more reproducible code in TransPolymer's github repository, we primarily evaluated the performance of the TransPolymer model on the three regression tasks in our work, while the data for polyBERT was sourced from its respective paper (both TransPolymer and polyBERT did not address any classification tasks in their study).

We selected the TransPolymer models pre-trained on 0.05M and 5M data. Additionally, we trained a version based on randomly sampled 0.02M data (under the parameters and settings provided by TransPolymer), taking into account the data volume used by PolyNC, to compare the performance of both models under an equal data volume due to the data-greedy nature of LLMs. Based on the experimental results (**Table S4**), we have demonstrated that when using R^2 as the evaluation metric and

considering an equal training data volume, PolyNC performed comparably to TransPolymer. As the training data of TransPolymer increased, its performance continued to improve. According to the results from polyBERT's paper, it has also been observed that the model's performance on the T_g task improves with larger data volumes. These findings regarding data volume suggest that PolyNC has the potential to further enhance its performance as the data volume increases, which will be a focus of our future efforts. It is worth noting that both PolyNC and polyBERT were trained on multiple tasks simultaneously, and the distribution shift between tasks can potentially be detrimental to the performance. For example, even with a larger training dataset, polyBERT performs worse than PolyNC and TransPolymer on the AE task. Indeed, despite the potential risk of performance degradation due to distributional shifts, the impressive end-to-end multitasking capability of PolyNC is still highly impressive.

Table S4. Performance comparison between PolyNC and recently models. The performance of TransPolymer and polyBERT is comparable. Due to the convenience of code and data availability, we chose to conduct comparative experiments using TransPolymer. The data for polyBERT was obtained from its paper (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-023-39868-6/MediaObjects/41467_2023_39868_MOESM1_ESM.pdf)

	PolyNC	**TransPolymer**			**polyBERT**
# Train data	0.02M	0.02M	0.05M	5M	100M
R^2--T_g	0.85	0.85	0.90	0.91	0.92±0.01
R^2--BC	0.85	0.86	0.89	0.90	-
R^2--AE	0.89	0.87	0.92	0.96	0.85±0.02
is *end-to-end* ?	True		False		False
is *multi-tasks* ?	True		False		True
is *multi-types* ?	True		True		False

It is worth acknowledging that TransPolymer takes into account factors such as molecular weight, temperature, and other conditions to improve the accuracy of property predictions for specific tasks. And the natural language prompts in PolyNC can also play a significant role in facilitating the description of these structure-independent features. This would require a larger amount of data for training, and we

are mining data that includes these non-structural factors to further enhance the functionality of PolyNC.

S7. The Sensitivity of PolyNC to SMILES Variations

After performing data augmentation based on SMILES enumeration, our dataset has been expanded. Different SMILES notations of the same polymer are distinct inputs for the language model. However, since they correspond to the same underlying structure, the model's predicted results should be consistent. To study the robustness of PolyNC to SMILES variations, we plotted the distributions of the average values and standard deviations of PolyNC for different SMILES representations of the same structure on the three regression tasks (T_g, BC and AE), which have stronger potential variability than the classification task. From **Figure S11**, it can be observed that PolyNC shows acceptable errors in predicting different SMILES for the same structure (after all, different SMILES representations can alter the order of tokens, thereby changing the actual input to the model).

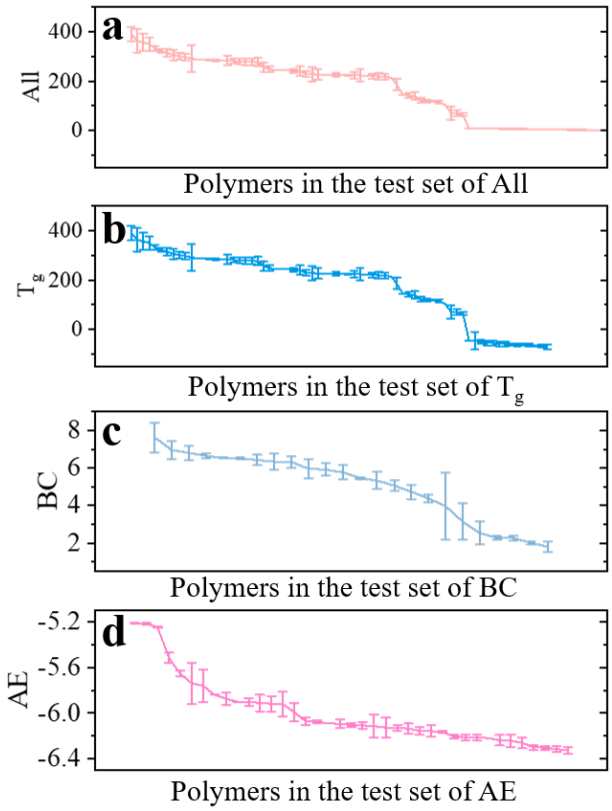
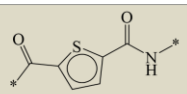


Figure S11. The distributions of the average values and standard deviations of PolyNC for different

SMILES representations of the same structure are shown across three regression tasks (**a**) and for T_g (**b**), BC (**c**), and AE (**d**). The x-axis represents different polymer structures, while the y-axis represents the average prediction values of different SMILES representations for each structure by PolyNC. The error bars represent the standard deviation of PolyNC across different SMILES representations of each structure.

We present a complex polymer structure (containing carbonyl, cyclic, amide, etc.) as a sample in **Table S5**, listing 15 different SMILES variations and model predictions for each SMILES variation, and find that PolyNC exhibits similar or equal and highly accurate predictions for the different SMILES variations of this structure.

Table S5. An example demonstration of a complex polymer structure (containing ester groups, rings, amide groups, etc.), PolyNC exhibits close to identical and highly accurate predictions for different SMILES representations of this structure. We have uploaded the results on all structures in the test sets at: https://github.com/HKQiu/Unified_ML4Polymers/blob/main/data/PromptAndPolyNCPrediction.csv.



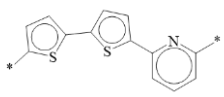
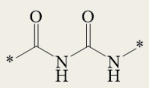
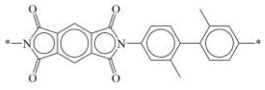
SMILES Variation	Gound-Truth	PolyNC's Prediction
c1c(C(N*)=O)sc(C(*)=O)c1	-6.4	-6.35
C(c1sc(C(=O)*)cc1)(N*)=O	-6.4	-6.22
c1(C(N*)=O)ccc(C(=O)*)s1	-6.4	-6.21
c1c(C(N*)=O)sc(C(=O)*)c1	-6.4	-6.35
C(c1sc(C(N*)=O)cc1)(=O)*	-6.4	-6.22
C(=O)(*)c1sc(C(N*)=O)cc1	-6.4	-6.22
O=C(*)c1ccc(C(N*)=O)s1	-6.4	-6.21
*NC(=O)c1ccc(C(*)=O)s1	-6.4	-6.22
O=C(N*)c1ccc(C(=O)*)s1	-6.4	-6.21
C(c1sc(C(=O)N)cc1)=O	-6.4	-6.22
*NC(c1sc(C(*)=O)cc1)=O	-6.4	-6.22
s1c(C(N*)=O)ccc1C(=O)*	-6.4	-6.21
C(c1ccc(C(N*)=O)s1)(=O)*	-6.4	-6.35
O=C(c1sc(C(=O)*)cc1)N*	-6.4	-6.22
O=C(c1ccc(C(*)=O)s1)N*	-6.4	-6.21
c1c(C(N*)=O)sc(C(*)=O)c1	-6.4	-6.35

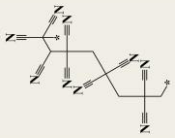
Above findings demonstrated that PolyNC's predictions depend more on the underlying structure rather than the different notations of SMILES, as reflected in the relatively small deviations in the predicted results for most molecules. But it needs to be emphasized that PolyNC needs further strengthening in recognizing different SMILES of the same structure to eliminate the deviations, and incorporating more training data could further enhance the model's robustness.

S8. Predictive Capability of PolyNC for Complex Molecular Structures

Complex structures can be found in organic molecules, such as cis-trans isomerism and chirality. These complex structures can be described using SMILES notation. For example, C2H2F2 can be represented as F/C=C\F (cis-) and F/C=C/F (trans-), and the "@ " and "@@" symbols are used to denote L and D chirality, respectively. The complex structures that can be represented using SMILES notation can all be used as inputs for PolyNC. However, our dataset contains limited information in this regard, making it necessary to acquire additional data in order to enhance its predictive capabilities for such structures. In addition, we also validated PolyNC's capability in handling typical complex structures in polymers: rings, homopolymers, copolymers, and branching. For cyclic polymers, we have selected a molecule with heterocyclic structures as an example. For branched polymers, we have chosen a structure with multiple side groups as an example. This selection is because the SMILES syntax remains consistent in handling these structures. The results are detailed in **Table S6.** These examples demonstrate PolyNC's proficiency in dealing with complex polymer structures. Furthermore, features related to polymer processing, such as number-average molecular weight (Mn), temperature, etc., which are beyond the scope of SMILES, can be described in the natural language prompt with more additional training data.

Table S6. Examples of complex polymer structure and PolyNC's predictions.

Type	Structure	Prompt	Pred. (Exp.)
rings		Predict the bandgap crystal of the following SMILES: c1(*)cccc(-c2ccc(-c3sc(*)cc3)s2)n1	2.32 (2.05[1]) eV
homo-polymers		Predict the atomization energy of the following SMILES: O=C(N*)NC(=O)*	-6.22 (-6.46[2]) eV
copolymers (alternating)		Predict the Tg of the following SMILES: c1cc(-c2c(C)cc(*)cc2)c(C)cc1-n1c	385 (374[3]) ℃

		(=O)c2cc3c(=O)n(*)c(=O)c3cc2c1=O	
branching		Predict the bandgap crystal of the following SMILES: C(C(CC(C(C(*)(C#N)C#N)C#N)(C#N)C#N)(CC(C#N)(C*)C#N)C#N)#N	6.45 (6.45[4]) eV

References:

(1) 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.

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

(3) 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 (8), 2930–2940.

(4) 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.