On-Demand Reverse Design of Polymers with PolyTAO

Haoke Qiu[1,2] and Zhao-Yan Sun[1,2*]

[1*]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, Jilin, China.

[2]School of Applied Chemistry and Engineering, University of Science and Technology of China, Hefei, 230026, Anhui, China.

*Corresponding author(s). E-mail(s): zysun@ciac.ac.cn;
Contributing authors: hkqiu@ciac.ac.cn;

Abstract

The forward screening and reverse design of drug molecules, inorganic molecules, and polymers with enhanced properties are vital for accelerating the transition from laboratory research to market application. Specifically, due to the scarcity of large-scale datasets, the discovery of polymers via materials informatics is particularly challenging. Nonetheless, scientists have developed various machine learning models for polymer structure-property relationships using only small polymer datasets, thereby advancing the forward screening process of polymers. However, the success of this approach ultimately depends on the diversity of the candidate pool, and exhaustively enumerating all possible polymer structures through human imagination is impractical. Consequently, achieving on-demand reverse design of polymers is essential. In this work, we curate an immense polymer dataset containing nearly one million polymeric structure-property pairs based on expert knowledge. Leveraging this dataset, we propose a Transformer-Assisted Oriented pretrained model for on-demand polymer generation (PolyTAO). This model produces polymers with 99.27% chemical validity in top-1 generation mode (approximately 200k generated polymers), representing the highest reported success rate among polymer generative models. Additionally, the average $\boldsymbol{R^2}$ between the properties of the generated polymers and their expected values across 15 predefined properties is 0.96. To further evaluate the pretrained model's performance in generating polymers with additional user-defined properties for downstream tasks, we conduct fine-tuning experiments on

https://doi.org/10.26434/chemrxiv-2024-3z7tw-v2 **ORCID:** https://orcid.org/0000-0003-4083-5507

three publicly available small polymer datasets using both semi-template and template-free generation paradigms. Through these extensive experiments, we demonstrate that our pretrained model and its fine-tuned versions are capable of achieving on-demand reverse design of polymers with specified properties, whether in semi-template generation or the more challenging template-free generation scenarios, showcasing its potential as a unified pretrained foundation model for polymer generation.

Keywords: Reverse On-Demand Design; Polymer Discovery; Generative Model; Large Language Model

1 Introduction

The array of potential materials available on Earth is staggering, with estimations reaching as high as 10^{60}[1]. However, this figure may prove to be even more expansive in reality, considering factors such as lattice defects in inorganic materials and the stochastic, multi-scale structures inherent in polymers[2].

Machine learning (ML) has emerged as a formidable tool in the quest for efficiently discovering candidate structures capable of serving as viable 'materials', showcasing notable accuracy and efficiency across various domains including inorganic materials[3, 4], metal materials[5, 6], organic molecules[7, 8], and polymer materials[9, 10]. Yet, the journey towards developing polymer materials presents distinct challenges owing to the scarcity of data and the intricate cross-scale structure-property relationships[11–16]. In the realm of ML-assisted polymeric materials discovery, two primary methodologies can be delineated: **Forward Screening** and **Reverse Design**. The Forward Screening entails the utilization of models to sift through candidate structures from a predetermined pool of potential polymers. A gamut of ML models, ranging from rudimentary to sophisticated, including feed forward neural network[17, 18], convolutional neural networks[19, 20], graph neural networks[21–23], recurrent neural networks[24], and more recently, Transformer models[16, 25, 26], have been used to establish surrogate models for polymer forward screening. Although this approach yields commendable efficacy, particularly with small polymer datasets[14, 27, 28], there remains a risk of overlooking structures that transcend human imagination[29, 30].

Conversely, the Reverse Design paradigm enables the direct, on-demand design of candidate structures tailored to meet performance specifications, obviating the necessity for a predetermined pool of candidates. This represents a more optimal strategy for bespoke polymer design and harbors the potential to yield candidate structures that elude expert intuition. At its core, this paradigm is underpinned by generative models, such as the variational autoencoder (VAE)[31, 32], diffusion models[33, 34], and Transformer[35, 36], and has witnessed groundbreaking advancements, particularly in the design of organic small molecules and drugs[31, 34, 37], with the percentage of chemically valid molecules generated surpassing 99%[34].

Inspired by the success of molecular generation models on small molecules and drugs, polymer scientists are also endeavoring to develop generative models tailored to the dynamic demands of polymer applications. Batra et al.[29] introduced a modified VAE designed to generate polymer repeat units based on SMILES notation, while they found the chemical validity of the generated polymers was less than 30%. This stark contrast with the higher validity observed in generative models for small organic molecules is largely due to the presence of two unique characters ('*') in polymer SMILES strings[15, 29]. These characters, which do not correspond to any chemical elements but signify distinct polymerization points[38], add complexity to polymer generation and diminish the performance of generative models trained on limited datasets[15]. Indeed, training on larger datasets holds promise for enabling the model to learn the intricacies of polymer chemistry[16, 25, 26]. Meanwhile, polymer scientists attempted to represent polymers using molecular graphs as an alternative representation method to enhance the chemical validity of generated polymers. Kim et al.[15], Liu et al.[39], and Gurnani et al.[30] have respectively employed graph neural networks for training polymer generative models. These efforts have yielded significant improvements in the chemical validity of generated polymers, with success rates ranging from 16.07 to 89.40%, 44.03%, and 93%, respectively.

Although there has been improvement in the proportion of chemically valid polymers generated at present, current polymer generative models still encounter significant challenges: 1) Foundational models can help boost the performance of downstream tasks[40], especially in the field of chemistry where the small data phenomenon is frequently encountered. The absence of pre-trained foundation models for polymer generation is a critical limitation. Due to constrained polymer datasets, polymer scientists often train polymer generative models from scratch using small, property-specific datasets, such as those tailored for dielectric performance [30, 39]. However, a pre-trained foundation model holds the promise of leveraging small datasets for various polymer properties, facilitating the accurate generation of polymers with diverse properties. 2) Current polymer generative models are trained based on SMILES-to-SMILES translation[29] or graph-to-graph translation[15, 30, 39] (or reconstruction). The unsupervised nature of this strategy inherently requires more data to learn hidden chemical patterns. As a result, the ability of the current polymer generative models to generate chemically valid polymers is limited. 3) When generating new molecules, these unsupervised approaches typically involves modifying the numerical representations within the hidden layers of generative models, which are then decoded into new polymer structures. Thus this approach necessitates an initial template polymer with the desired properties, yet identifying such templates is a formidable challenge[41]. Besides, due to the large dimensionality of the editable numeric representation of the template polymer, the directionality of reverse design is partially out of control[31, 39], introducing uncertainty to the screening task. 4) Additionally, this generation often occurs within the neighborhood of the template polymer[39], and it is difficult to efficiently explore the diverse polymer space as polymers with low structural similarity may exhibit similar properties. For instance, a polymer chain containing hydrogen bonding interactions may exhibit a similar glass transition temperature to another polymer chain containing multiple benzene rings[41].

To address these challenges, we refined a polymer structure-property dataset containing nearly 1,000,000 entries based on the largest unlabeled polymer dataset, PI1M[42]. Using this curated structure-property dataset, we propose PolyTAO, a polymer generative pre-trained large language model (LLM), via supervised learning (Figure 1). The pre-trained model demonstrates impressive polymer generation capabilities, with chemical validity exceeding 99% when generating a total of approximately 200,000 polymers in top-1 mode. We calculated the 15 pre-defined fundamental properties of the generated polymers, showing extremely high prediction accuracy with the expected values (the average R^2 is 0.96). We further tested its ability to generate polymers with other user-defined properties in multiple downstream tasks and explored the feasibility of the progressive semi-template and completely template-free polymer generation. The results demonstrate its excellent performance in on-demand reverse polymer generation, and the generated polymers exhibits diverse structural features, which showcases the model's ability to thoroughly explore the polymer space, as well as its capabilities as a foundational polymer generative model.

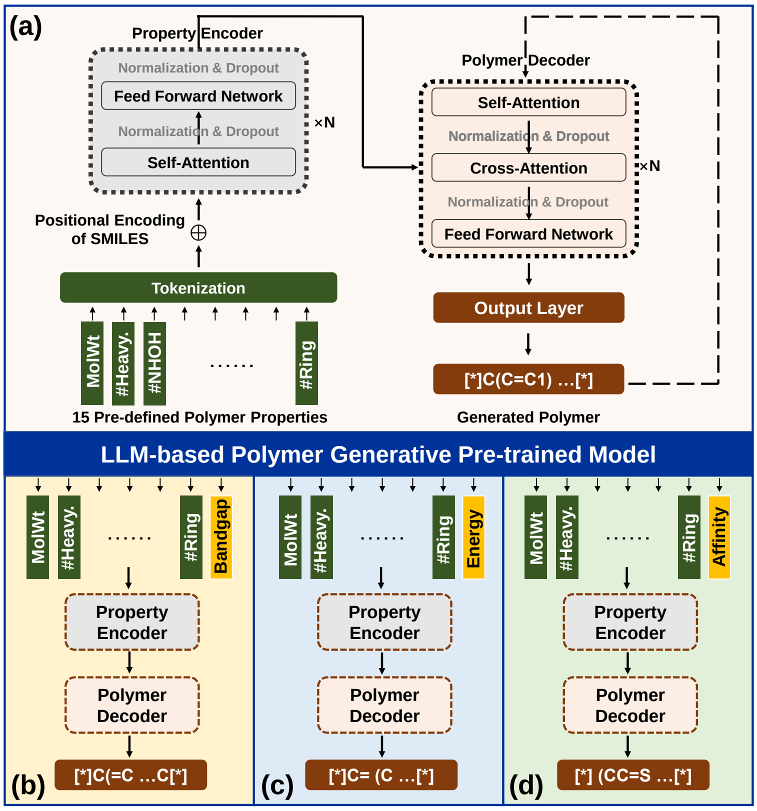

Fig. 1 (a) Model architecture of PolyTAO. Using a set of 15 predefined fundamental features related to polymers as input, it has demonstrated impressive accuracy in generating the SMILES of polymer repeat units that satisfy these features via supervised learning for conditional generation. Subsequently, we validated the model's ability to generate polymers with other user-defined properties by generating polymers with specified band gaps (b), atomization energy(c), and electron affinity(d).

2 Results

2.1 Performance of PolyTAO on polymer generation

2.1.1 Top-1 generation

We first conducted top-1 generation experiments (i.e., generating one polymer for each input) using the pre-trained model on the test set, resulting in a total of 199,159 samples generated. Among them, 99.27% were chemically valid, which represents the highest value among existing polymer generative models to date on the largest test sample (**Table 1**). This demonstrates that the model has deeply learned the mapping between polymeric fundamental properties and SMILES after pretraining via large-scale supervised learning.

Table 1 Performance of the pre-trained model on the test set via top-1 generation. # Data: number of training data; # Gen.: number of generated polymers; Val.: validity; UNC: unconditional; CND: conditional. [a] Results for two polymer properties (glass transition temperature and band gap). [b] Results for one polymer property (i.e., the partition coefficient, logP). [c] Results for 15 polymer properties (as illustrated in Methods).

Model	Architecture	Mode	# Data	# Gen.	Val./% ↑	Average R^2↑
SD-VAE[29]	CNN	UNC	250k	1k	13-27	0.65[a]
polyG2G[30]	GNN	UNC	13k	58k	93	
IGGM[39]	GNN	UNC	250k	10k	44.03	
Mole. Chef[15]	GNN	UNC	120k		16.07-89.40	0.96[b]
Ours	**Transformer**	**CND**	**800k**	**199k**	**99.27**	**0.96[c]**

We conducted a statistical analysis of the types of chemical elements present in our polymer structure-property dataset and the generated polymers (Figure 2). The most abundant elements are C, N, and O, followed by other inorganic elements such as S and F, while metal elements constitute a smaller proportion, which aligns with the empirical knowledge in polymer science. The distribution of element proportions in the training set (Figure 2a) is similar to that in the test set (Figure 2b), indicating a relatively uniform dataset partition. Interestingly, in the polymers generated using the top-1 mode, some metal elements are not generated (Figure 2c). This is because that each LLM generates tokens (i.e., chemical elements) based on the probability of each token's occurrence, and tokens with very low probabilities may not be generated. If necessary, this can be optimized by increasing the proportion of metal polymers in downstream tasks.

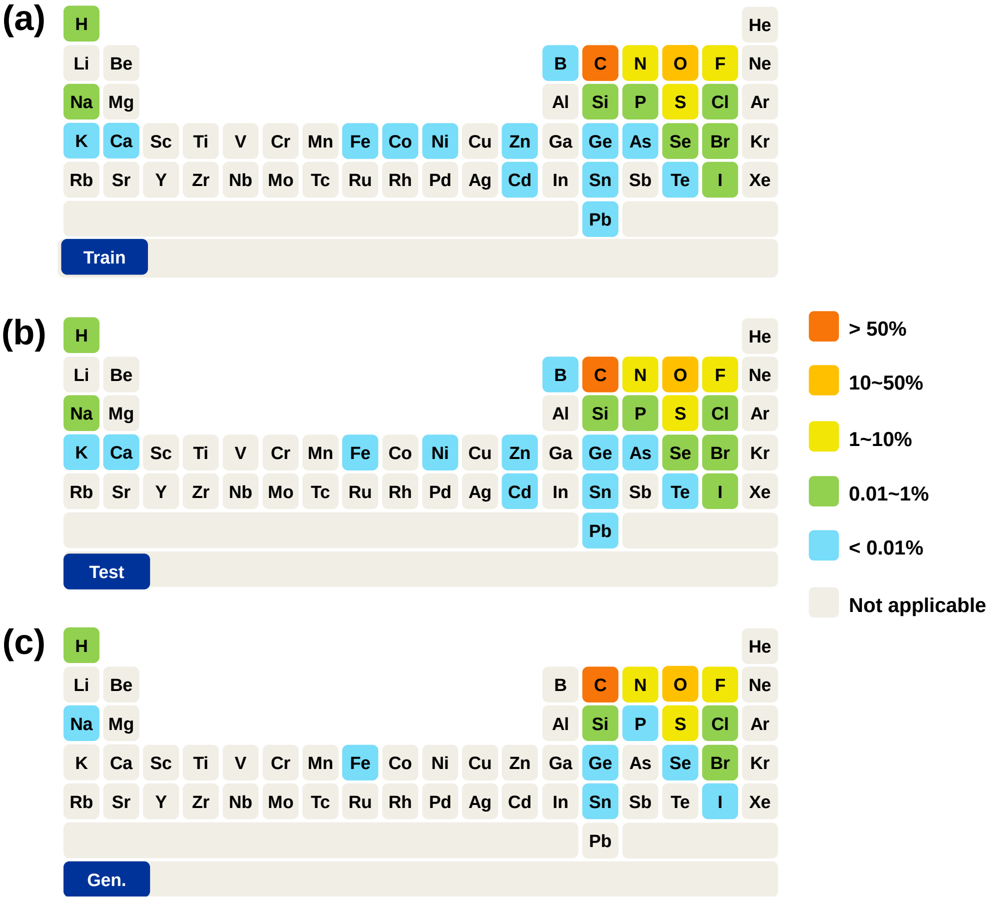

Fig. 2 Statistical analysis of the types of chemical elements of the: training set (a), test set (b) and generated polymers (c). Gen.: Generated.

In order to demonstrate whether the generated polymers possess the expected fundamental properties specified in the input parameters, we examined the aforementioned properties of the generated polymers, where we found a high degree of agreement between them (Figure 3). For the chemically valid and unique polymers generated by the pre-trained model, the average R^2 value across the 15 polymer properties is 0.96. This indicates that PolyTAO can preliminarily achieve the on-demand generation of polymers with specified properties. This result instills confidence in our model's ability to generate polymers with other properties, as discussed in the *Applications of PolyTAO in generating polymers with other user-defined properties* section. Interestingly, for certain properties, such as the number of aliphatic carboncycles (NumAliphaticCarboncycles) and the number of aliphatic heterocycles (NumAliphaticHeterocycles), the generated polymers exhibit relatively poor consistency. However, these inconsistencies are advantageous for the model to produce structurally diverse polymers, facilitating the generation of various ring structures (as illustrated in **S2** of SI).

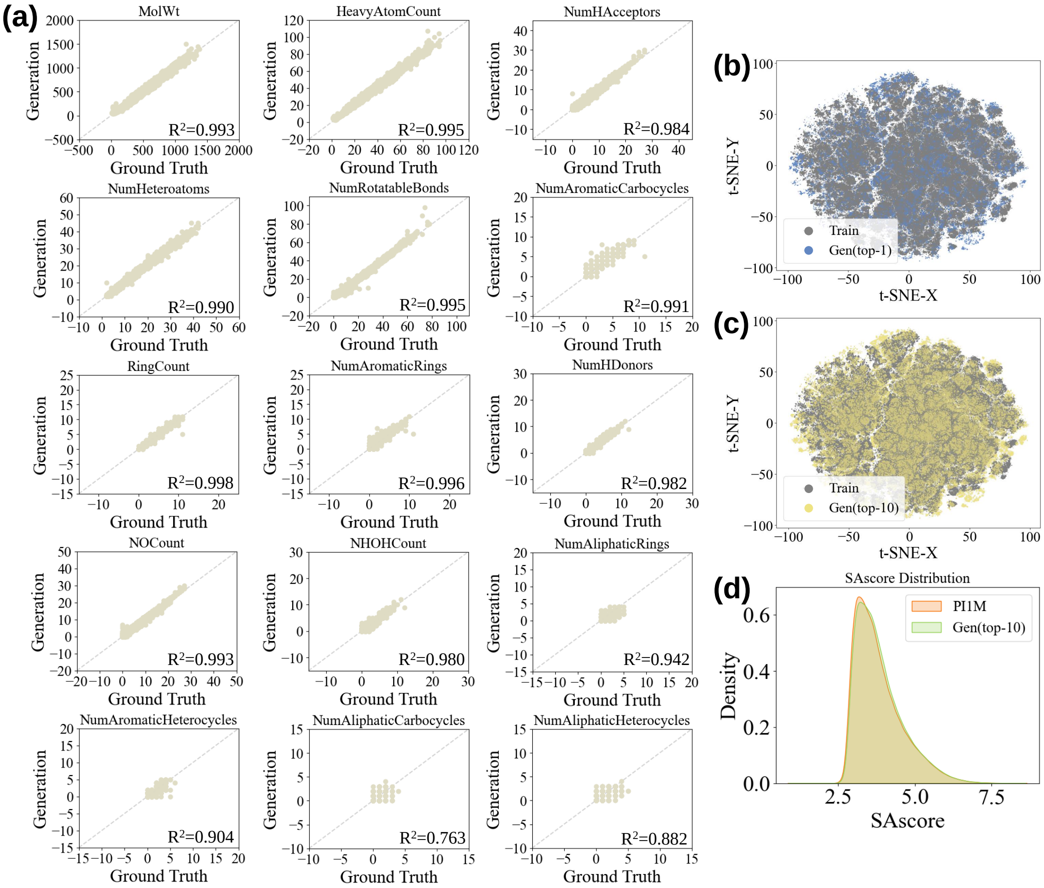

Fig. 3 The fitting plots of the properties generated by the model (Generation) against the expected properties (Ground Truth). (b)-(c) The chemical space of polymers after t-SNE dimensionality reduction, with the background color indicating the randomly selected training set from PI1M and (b) represents top-1 generation, with (c) representing top-10 generation. (d) illustrates the SAscore of PI1M and the valid, unique, and novel molecules generated under the top-10 mode.

2.1.2 Top-k (k>1) generation

Due to the stochastic and probabilistic nature of generation of LLMs, to validate the generation stability of PolyTAO, we examined the model's performance in top-k generation. Specifically, we generated three (top-3), five (top-5), and ten (top-10) samples for the same input, evaluating the Validity, Uniqueness, and Novelty of the generated polymers. Additionally, we assessed the Tanimoto similarity coefficient between the generated polymers and input samples, as well as the synthesizability (SAscore) of the generated polymers. The results are reflected as **Table 2**. In the top-k generation mode, the model tends to be "adventurous" in its generation, resulting in polymers that maintain a high level of uniqueness and novelty. Compared to top-1 generation (Figure 3b), top-k generation expands the chemical space of generated polymers, even extending beyond the chemical space corresponding to the training set (Figure 3c). Regardless of the value of k, the chemical similarity between the generated polymer and the polymer corresponding to the input prompt is consistently low. In contrast, generating new molecules based on artificially modified latent representations of molecules tend to produce molecules with very high similarity[39]. However, this "adventurous" generation can also lead to the generation of chemically invalid polymers, resulting in a slight decrease in the chemical validity of the generated polymers compared to top-1 generation, but still higher than previous polymer generative models. Impressively, for all valid, unique and novel molecules generated under the top-10 mode (totaling 1,828,027), and their synthesizability did not become more challenging, demonstrating a synthesizability similar to that of PI1M as illustrated in Figure 3d.

Table 2 Performance of the pre-trained model on the test set via top-k (k=3,5,10) generation.

Metric	Top-3	Top-5	Top-10
Validity ↑	97.75±0.0001	97.76±0.0002	97.75±0.0004
Uniqueness ↑	99.07±0.0001	99.06±0.0002	99.08±0.0001
Novelty ↑	93.56±0.0009	93.73±0.0005	94.01±0.0006
Similarity ↓	0.302±0.0002	0.303±0.0002	0.306±0.0002
SAscore ↓	3.84±0.77	3.83±0.77	3.85±0.77

2.2 Applications of PolyTAO in generating polymers with other user-defined properties

The above results demonstrate the impressive capability of PolyTAO in generating polymers with pre-defined foundational properties. Then we further assessed the model's ability to generate polymers with other user-defined properties to demonstrate its robustness and universality in the on-demand design of polymers.

In principle, polymer generation based on SMILES-to-SMILES translation or graph-to-graph translation require a template polymer that meets the desired performance[15, 29, 30, 39]. By editing the latent representation of the template polymer and decoding this representation, new polymers can be generated. We refer to this paradigm as **template-based** polymer generation. However, due to the randomness in editing the latent representation, achieving on-demand design through this method is limited[39]. Additionally, finding template polymers that meet the requirements is also an arduous and challenging task[38, 41, 43–45]. Instinctively, By incorporating the target properties directly into the input prompts in a similar manner to the pre-training phase, there is potential to achieve on-demand generation of polymers without providing template polymers. Though this approach represents an advancement compared to template-based polymer generation, the input at this stage includes not only the target properties but also the 15 fundamental polymer properties we defined. Therefore we define this paradigm as **semi-template** generation.

Then, we tested the performance of the pre-trained model in the semi-template generation scenario. We finetuned this pre-trained model on ten public polymers datasets of different properties (**S3** of SI) to obtain expert LLMs for each property. However, since experimentally validating the properties of generated polymers on a large scale is resource-intensive, we attempted to train proxy models for each dataset to efficiently validate the properties of the generated polymers on large-scale. To be specific, we utilized graph neural networks, known for their excellent performance in molecular property prediction, to train proxy models for each property. Since the accuracy, i.e., coefficient of determination (R^2), of the proxy models is crucial in assessing the properties of the generated polymers and the performance of expert LLMs, we selected the top three proxy models with the highest R^2 (exceeding 0.9, detailed in **S3** of SI) and the corresponding dataset as subsequent case studies. These proxy models are tailored for the following polymer properties: band gap, atomization energy, and electron affinity. We then finetuned the pre-trained model using these datasets individually, with the same data partitioning as when training the respective proxy models. Each polymer property served as an additional vector added to the input prompt (see Figure 1 (b)-(d)). After fine-tuning, the loss of the expert LLMs exhibited convergence (Figure 4a, 5a and 6a).

In the generation of expert LLMs, we conducted top-5 generation (i.e., generating five samples for each input) and repeated the process for three rounds, to mitigate the randomness of LLM and assess the model's ability to generate multiple polymers satisfying the target properties. Here, our main focus is to investigate the feasibility and reliability of LLM-based polymer generative model in practical usage. Therefore, we do not delve into detailed discussions regarding the uniqueness of generated polymers in the following sections.

We selected samples from the test sets for each property, following the criterion that the proxy model's prediction for the sample closely matches the ground truth, aiming to enhance the rationality of utilizing the proxy model to assess the properties of generated polymers (i.e., at least, the proxy model should be sufficiently accurate in predicting the properties of input samples). Moreover, we also aimed to ensure that the selected samples exhibit outstanding properties whenever possible.

Band gap The band gap of polymer holds significant importance in the advancement of polymer-based electronic and photonic devices, as it profoundly impacts their functionality across domains such as organic photovoltaics and light-emitting diodes. We take the example of generating polymers with wide band gaps (greater than 6 eV)[30]. Following the aforementioned selection criteria, we opted for the structure depicted in Figure 4a as the input sample, whose band gap is 6.23 eV measured by the proxy model, with its 15 fundamental properties plus its band gap serves as the input prompt. Across 3 rounds of top-5 generation, the expert LLM on band gap yielded 14 out of 15 chemically valid polymers, with 13 out of 14 showcasing novel structures from the input sample. Impressively, the predictions of the proxy model showcased that these 13 novel samples demonstrate properties that align with the target band gap (with a margin of error of 5% from 6.23).

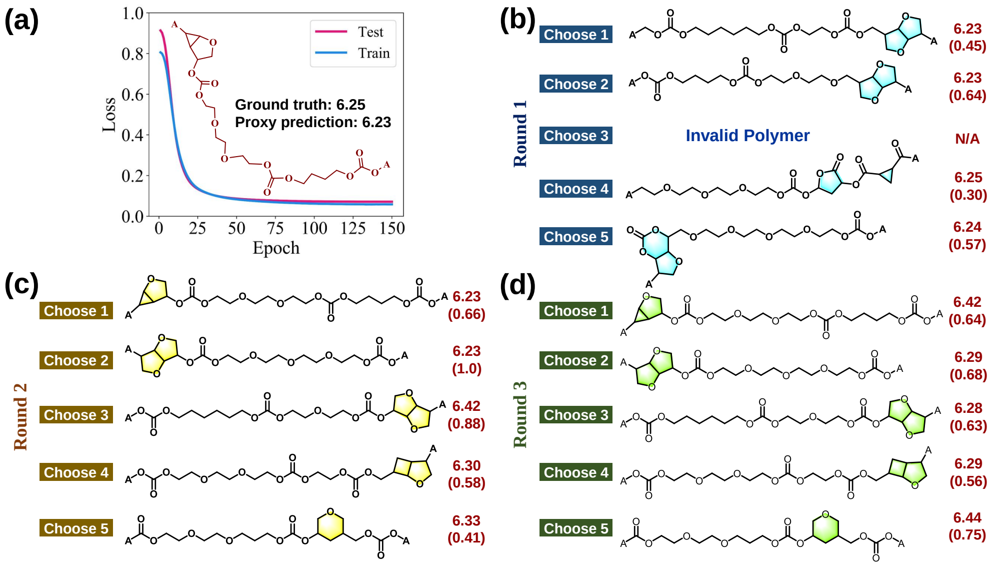

Fig. 4 Loss of the expert LLM on band gap and the input sample (a) and on-demand inverse generation on band gap task (b)-(d). The numbers in subfigures (b)-(d) represent the band gaps predicted by the proxy model (relative to the similarity with the input sample), i.e., *predicted band gap(similarity)*. During the fine-tuning of this task, the training set consisted of 3042 samples, while the test set comprised 338 samples.

Atomization Energy The atomization energy of polymers reflects the strength and stability of the bonds within polymer molecules. Similarly, from the test set of the atomization energy database, we selected a polymer with high atomization energy as the input sample (Figure 5a) and its atomization energy is -6.18 eV measured by the proxy model. The 15 fundamental properties plus the atomization energy of this polymer are used as the input prompt. After three rounds of top-5 generation, the expert LLM on atomization energy generated 100% chemically valid and novel polymers. It is noteworthy that the data size of this dataset is too small for a LLM, resulting in a slight decrease in the accuracy of generated polymers. According to the predictions of the proxy model, 11 out of 15 polymers exhibit properties that align with the target atomization energy (with a margin of error of 5% from -6.18).

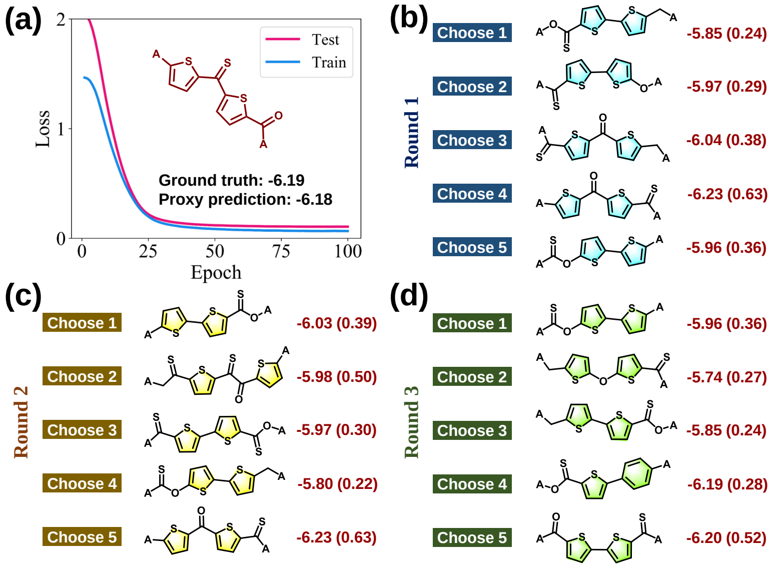

Fig. 5 Loss of the expert LLM on atomization energy and the input sample (a) and on-demand inverse generation on atomization energy task (b)-(d). The numbers in subfigures (b)-(d) represent the band gaps predicted by the proxy model (relative to the similarity with the input sample), i.e., *predicted atomization energy(similarity)*. During the fine-tuning of this task, the training set consisted of 351 samples, while the test set comprised 39 samples.

Electron Affinity The electron affinity of polymers reflects the polymer molecule's ability to accept electrons, a property crucial in photovoltaic applications and other electronic applications of polymers. Unfortunately, the dataset for this property is also relatively small. During the generation, we chose the 15 fundamental properties plus the electron affinity of the structure depicted in Figure 6a as the input prompt. The electron affinity of this polymer is 3.14 (measured by the proxy model). From the results of three rounds of top-5 generation, 100% of the generated polymers were chemically valid, with 14 out of 15 being novel. According to predictions from the proxy model, 10 out of 14 polymers exhibit properties that align with the target electron affinity (with a margin of error of 5% from 3.14).

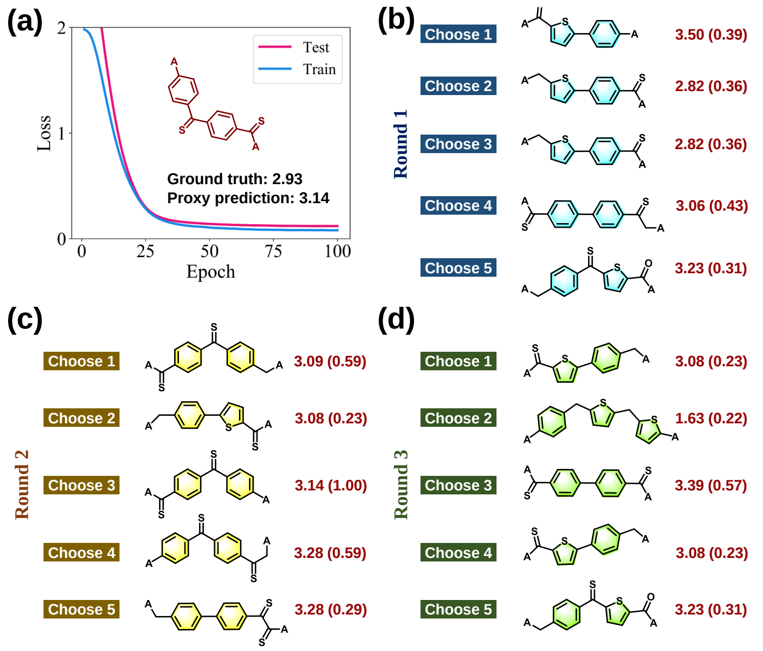

Fig. 6 Loss of the expert LLM on electron affinity and the input sample (a) and on-demand inverse generation on electron affinity task (b)-(d). The numbers in subfigures (b)-(d) represent the band gaps predicted by the proxy model (relative to the similarity with the input sample), i.e., *predicted electron affinity(similarity)*. During the fine-tuning of this task, the training set consisted of 331 samples, while the test set comprised 37 samples.

2.3 Template-free generation: an ambitious task towards on-demand polymer design

Compared to previous template-based approaches[15, 29, 30, 39], the semi-template method introduced above takes target properties as part of input, enabling the generation of polymers with specified properties. This represents an advancement over entirely template-based polymer generation. However, one more challenging goal is to achieve template-free polymer generation. The potential scenario for this paradigm is to provide the generative model with only a desired property value, allowing the model to freely generate structures that meet the requirements. Clearly, this design paradigm is ambitious yet more challenging due to the contradiction between the infinite chemical space of polymers and the limited training data available. To assess the feasibility of our pre-trained model in this challenging task, we conducted fine-tuning tests using the band gap dataset due to its larger number of data entries. During the implement, we utilized only the value of band gap as input, with the corresponding polymer repeat unit SMILES as output (Figure 7a).

Meanwhile, in order to simultaneously achieve a meaningful objective, our aim is to have the fine-tuned expert LLM generate polymers with higher band gaps. We used 6.5 eV (higher than previous semi-template generation task) as target, then the model was expected to generate polymer structures with band gaps around 6.5 eV. Throughout the generation process, we continued to utilize the top-5 generation mode and repeated the process for three rounds to assess the stability of the LLM-based polymer generative model. The generation results in Figure 7b indicate that the proportion of chemically valid molecules generated is 100%, which is a significant prerequisite for the success of this task. Furthermore, in this novel task previously unexplored by polymer scientists, as verified by the proxy model, the expert LLM can produce no fewer than 2 samples with target band gap (with a margin of error of 5%) in each round of generation (marked in red font in Figure 7b). In total, 9 polymers exhibit properties that align with the target band gap (with a margin of error of 5%). Interestingly, compared to template-based and semi-template-based polymer generation methods mentioned earlier, the template-free approach generates a more diverse range of polymer structures, showcasing the model's freedom to explore the polymer space. These results demonstrate the feasibility of on-demand reverse polymer generation without template.

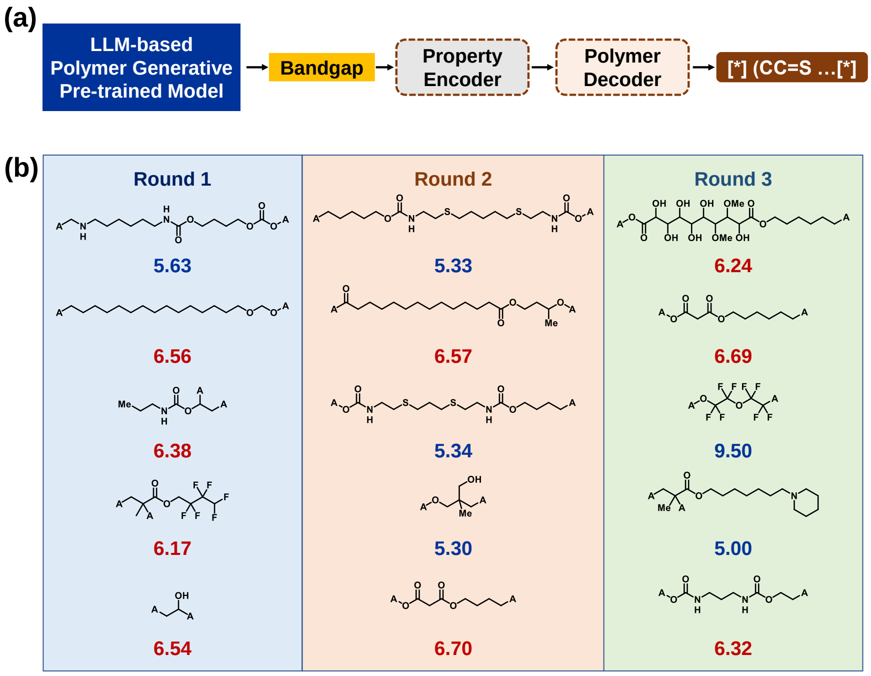

Fig. 7 (a) Fine-tuning the pre-trained model for the template-free generation. (b) The generation results via template-free generation. The model will aim to generate polymer structures that meet the specified target values.

3 Discussion and Conclusion

In this work, we proposed a generative pre-trained model based on LLMs for on-demand reverse polymer design. The pre-trained model was trained on a meticulously curated dataset containing nearly one million polymer structures and fundamental properties, crafted based on expert knowledge. Evaluation on a test set of nearly 200,000 samples revealed that the model generated chemically valid molecules with a proportion of 99.27%. Through further top-10 generation, the pre-trained model designed over 1.8 million valid and novel polymer structures, effectively doubling the entries of the off-the-shelf polymer datasets. These data, along with widely known datasets like PI1M, can offer a richer candidate pool for paradigms based on forward screening. To achieve the generation of polymers with other specific properties, we fine-tuned the pre-trained model on three publicly available polymer property datasets, resulting in expert LLMs tailored to each property. The generation results of these expert LLMs demonstrate the powerful capability of the model in on-demand reverse generation. However, for more precise on-demand design, we advocate for greater efforts from the polymer community to expand polymer property datasets. Additionally, we attempted an ambitious task using the dataset with a relatively large amount of data on band gap, aiming for completely template-free polymer generation. The results indicate that the fine-tuned expert LLM can achieve on-demand reverse polymer generation based solely on the provided values of the desired polymer properties.

In summary, we have demonstrated a pre-trained model for on-demand reverse generation of polymers, and its performance on multiple downstream datasets indicates its broad applicability and transferability. Meanwhile, by employing more advanced polymer representations, such as BigSMILES[46], coupled with a larger amount of polymer data, there is potential to further enhance the model's performance in on-demand polymer generation. Also, more efforts, including but not limited to advancing the acquisition of large-scale BigSMILES strings and collecting multimodal, multiscale polymer data, need to be put into practice.

4 Methods

4.1 Polymer structure-property dataset

The largest publicly available polymer structure-property dataset currently is PolyInfo, containing around 20,000 polymer structure-property pairs. However, this dataset is insufficient for training a LLM. Recently, a virtual polymer database, PI1M[42], has been extended from PolyInfo, comprising nearly one million polymer structures but lacking corresponding property values. Researchers have utilized PI1M for unsupervised pretraining for polymer generation but this unsupervised pretraining paradigm results in the limited capacity of generating chemically valid polymers[39]. To achieve unprecedented large-scale supervised learning on this largest polymer dataset, we opted to compute foundational properties for each polymer structure in PI1M as descriptors. Due to the significant influence of molecular interactions and chain structure on the properties of polymers at the microscopic level, we carefully selected 15 descriptors related to the factors mentioned above. Specifically, we considered: 1) Molecular weight, providing an approximate constraint on the number and types of atoms in the repeat unit. 2) Hydrogen bonds, including the types and quantities of hydrogen bond donors and acceptors. 3) Atom types, including the number of heteroatoms apart from the common carbon and hydrogen atoms in polymers. 4) Chain structure, including the types and quantities of rings and the number of rotatable bonds, which account for the flexibility of polymer molecules. For the specific list and the corresponding description, please refer to Section 1 (**S1**) of the Supporting Information (SI). The above foundational properties and their corresponding SMILES constitute our polymer structure-property dataset.

4.2 Prompt Engineering

Like any LLM, designing high-quality prompts is crucial for on-demand generation by the model. We computed 15 fundamental physicochemical properties for each polymer repeat unit's SMILES and concatenated them to form the input prompts for PolyTAO. Except for molecular weight, all other physicochemical properties are of integer type. To reduce the input token size and improve training efficiency, we also converted the molecular weight to an integer type. In fact, this approach, imprecisely specifying the molecular weight and instead slightly ”fuzzifying” it, increases the model's freedom when generating new molecules, thus facilitating the generation of structurally diverse molecules (as shown in **S2** of SI).

4.3 Model Settings

Currently, there are many open-source large language models (LLMs) available for pretraining in chemical tasks. However, our previous research demonstrates that LLM based on a deep understanding of chemical knowledge may perform better even on less data[26]. Taking into account both model complexity and computational device requirements, we have chosen our previously developed PolyNC[26] as the foundational model. PolyNC is a LLM based on polymer structures with over 22 million parameters, capable of predicting various properties such as the glass transition temperature of polymers, benefiting from the cross-attention mechanism[47].

4.4 Polymer generative pretraining via large-scale supervised learning

During the pretraining, the foundational properties will be concatenated as input, while the corresponding structures (SMILES) serve as the output (Figure 1(a)). Compared to generative models using SMILES-to-SMILES translation and graph-to-graph translation, our paradigm of property-to-SMILES aims to enable the model to capture more foundational properties of polymers and their corresponding structures (SMILES). We randomly partitioned the polymer structure-property dataset into a training set (80%, ∼0.8 million) and a test set (20%, ∼0.2 million) for pretraining. During fine-tuning for ”semi-template” generation with other user-defined properties, we added other polymer's properties as additional vectors to the input prompt (Figure 1(b)-(d)).

4.5 Model Metrics

For molecular generative models, the primary metric of interest is the chemical **Validity** (the percentage of chemically valid molecules), which is our foremost consideration. Additionally, LLM can produce multiple outputs for the same input (i.e., top-k generation), which is beneficial for generating structurally diverse candidate molecules and helps assess the stability of the model's generation capability. Therefore, for top-k generation scenarios, we also evaluate the **Uniqueness** (the percentage of chemically valid molecules generated that are mutually unique in each generation of the k times generation) and **Novelty** (the percentage of generated valid molecules not in the training set and the test set in each generation of the k times generation) of the generated polymers.

It is worth mentioning that previous polymer generative models rarely discussed the similarity of generated polymers to existing polymers and, more importantly, the synthetic feasibility of the generated polymers. Thus, in top-k generation scenarios, we additionally assess the **Similarity** (i.e., Tanimoto similarity[48]) and synthetic feasibility[49] (**SAscore**) of the generated polymers.

Data availability. The PI1M dataset is publicly available at https://github.com/RUIMINMA1996/PI1M. The 15 fundamental properties in the pre-training stage were calculated using RDKit package (version: 2023.3.2). Our pre-trained model is publicly available at https://huggingface.co/hkqiu/PolymerGenerationPretrainedModel. Any other data and code related to reproducing the results will be provided promptly upon request.

Code availability. The source codes of demos for generation polymers via semi-template and template-free are available at https://github.com/hkqiu/PolymerGenerationPretrainedModel.

Supplementary information. The following files are available free of charge.

- Polymer physicochemical properties selected
- Structurally diverse polymers generation
- Proxy model for evaluation the properties of the generated polymers

Acknowledgments. 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**). We sincerely thank Prof. Rampi Ramprasad, Prof. Ryo Yoshida, Prof. Ying Li, Prof. Zhi-Min Dang, Prof. Nicholas E. Jackson, Prof. Arash Khajeh and their colleagues for their pioneering works in the field of polymer generation models. We also thank to Ms. Jiatong Lyu from Jilin University for her creativity and inspiration to our work. The work is also supported by the hardware in the Network and Computing Center in Changchun Institute of Applied Chemistry, Chinese Academy of Sciences.

Declarations

The authors declare no competing interests.

References

[1] Sanchez-Lengeling, B., Aspuru-Guzik, A.: Inverse molecular design using machine learning: Generative models for matter engineering. Science **361**(6400), 360–365 (2018) https://doi.org/10.1126/science.aat2663

[2] Gormley, A.J., Webb, M.A.: Machine learning in combinatorial polymer chemistry. Nat. Rev. Mater. **6**(8), 642–644 (2021) https://doi.org/10.1038/s41578-021-00282-3

[3] Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., Cubuk, E.D.: Scaling deep learning for materials discovery. Nature **624**, 80–85 (2023)

[4] Szymanski, N.J., Rendy, B., Fei, Y., Kumar, R.E., He, T., Milsted, D., McDermott, M.J., Gallant, M., Cubuk, E.D., Merchant, A., Kim, H., Jain, A., Bartel, C.J., Persson, K., Zeng, Y., Ceder, G.: An autonomous laboratory for the accelerated synthesis of novel materials. Nature **624**, 86–91 (2023)

[5] Pétuya, R., Durdy, S., Antypov, D., Gaultois, M.W., Berry, N.G., Darling, G.R., Katsoulidis, A.P., Dyer, M.S., Rosseinsky, M.J.: Machine-learning prediction of metal–organic framework guest accessibility from linker and metal chemistry. Angew. Chem. Int. Ed. **61**(9), 202114573 (2022) https://doi.org/10.1002/anie.202114573 https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202114573

[6] Nandy, A., Duan, C., Taylor, M.G., Liu, F., Steeves, A.H., Kulik, H.J.: Computational discovery of transition-metal complexes: From high-throughput screening to machine learning. Chem. Rev. **121**(16), 9927–10000 (2021)

[7] Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R.S., Cooper, A.I.: A mobile robotic chemist. Nature **583**(7815), 237–241 (2020) https://doi.org/10.1038/s41586-020-2442-2

[8] Li, X., Maffettone, P.M., Che, Y., Liu, T., Chen, L., Cooper, A.I.: Combining machine learning and high-throughput experimentation to discover photocatalytically active organic molecules. Chem. Sci. **12**, 10742–10754 (2021) https://doi.org/10.1039/D1SC02150H

[9] Lu, H., Diaz, D.J., Czarnecki, N.J., Zhu, C., Kim, W., Shroff, R., Acosta, D.J., Alexander, B.R., Cole, H.O., Zhang, Y., Lynd, N.A., Ellington, A.D., Alper, H.S.: Machine learning-aided engineering of hydrolases for PET depolymerization. Nature **604**(7907), 662–667 (2022)

[10] 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. **14**(1), 4838 (2023) https://doi.org/10.1038/s41467-023-40459-8

[11] Wu, S., Kondo, Y., Kakimoto, M.-a., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J., Yoshida, R.: Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput. Mater. **5**(1), 66 (2019) https://doi.org/10.1038/s41524-019-0203-2

[12] Kuenneth, C., Rajan, A.C., Tran, H., Chen, L., Kim, C., Ramprasad, R.: Polymer informatics with multi-task learning. Patterns **2**(4), 100238 (2021) https://doi.org/10.1016/j.patter.2021.100238

[13] 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. **8**(29), 9545 (2022)

[14] Barnett, J.W., Bilchak, C.R., Wang, Y., Benicewicz, B.C., Murdock, L.A., Bereau, T., Sanat K. Kumar: Designing exceptional gas-separation polymer membranes using machine learning. Sci. Adv. **6**(20), 4301 (2020) https://doi.org/10.1126/sciadv.aaz4301 https://www.science.org/doi/pdf/10.1126/sciadv.aaz4301

[15] Kim, S., Schroeder, C.M., Jackson, N.E.: Open macromolecular genome: Generative design of synthetically accessible polymers. ACS Polymers Au **3**(4), 318–330 (2023) https://doi.org/10.1021/acspolymersau.3c00003

[16] Xu, C., Wang, Y., Barati Farimani, A.: TransPolymer: A Transformer-based language model for polymer property predictions. npj Comput. Mater. **9**(1), 64 (2023) https://doi.org/10.1038/s41524-023-01016-5

[17] 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 **256**, 125216 (2022) https://doi.org/10.1016/j.polymer.2022.125216

[18] Zhang, S., He, X., Xia, X., Xiao, P., Wu, Q., Zheng, F., Lu, Q.: Machine-Learning-Enabled Framework in Engineering Plastics Discovery: A Case Study of Designing Polyimides with Desired Glass-Transition Temperature. ACS Appl. Mater. Interfaces **15**(31), 37893–37902 (2023) https://doi.org/10.1021/acsami.3c05376

[19] Tao, L., Chen, G., Li, Y.: Machine learning discovery of high-temperature polymers. Patterns **2**(4) (2021) https://doi.org/10.1016/j.patter.2021.100225

[20] Yan, C., Feng, X., Li, G.: From drug molecules to thermoset shape memory polymers: A machine learning approach. ACS Appl. Mater. Interfaces **13**(50), 60508–60521 (2021) https://doi.org/10.1021/acsami.1c20947

[21] Aldeghi, M., Coley, C.W.: A graph representation of molecular ensembles for polymer property prediction. Chem. Sci. **13**(35), 10486–10498 (2022) https://doi.

org/10.1039/D2SC02839E

[22] 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. **9**(1), 90 (2023) https://doi.org/10.1038/s41524-023-01034-3

[23] Wang, M., Jiang, J.: Accelerating Discovery of Polyimides with Intrinsic Microporosity for Membrane-Based Gas Separation: Synergizing Physics-Informed Performance Metrics and Active Learning. Adv. Funct. Mater., 2314683 (2024) https://doi.org/10.1002/adfm.202314683

[24] Simine, L., Allen, T.C., Rossky, P.J.: Predicting optical spectra for optoelectronic polymers using coarse-grained models and recurrent neural networks. Proc. Natl. Acad. Sci. **117**(25), 13945–13948 (2020) https://doi.org/10.1073/pnas.1918696117

[25] Kuenneth, C., Ramprasad, R.: polyBERT: A chemical language model to enable fully machine-driven ultrafast polymer informatics. Nat. Commun. **14**(1), 4099 (2023) https://doi.org/10.1038/s41467-023-39868-6

[26] 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. **15**(2), 534–544 (2024) https://doi.org/10.1039/D3SC05079C

[27] Bradford, G., Lopez, J., Ruza, J., Stolberg, M.A., Osterude, R., Johnson, J.A., Gomez-Bombarelli, R., Shao-Horn, Y.: Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery. ACS Cent. Sci. **9**(2), 206–216 (2023) https://doi.org/10.1021/acscentsci.2c01123

[28] 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 **11**(8), 2930–2940 (2023) https://doi.org/10.1039/D2TC05174E

[29] Batra, R., Dai, H., Huan, T.D., Chen, L., Kim, C., Gutekunst, W.R., Song, L., Ramprasad, R.: Polymers for Extreme Conditions Designed Using Syntax-Directed Variational Autoencoders. Chem. Mater. **32**(24), 10489–10500 (2020) https://doi.org/10.1021/acs.chemmater.0c03332

[30] Gurnani, R., Kamal, D., Tran, H., Sahu, H., Scharm, K., Ashraf, U., Ramprasad, R.: polyg2g: A novel machine learning algorithm applied to the generative design of polymer dielectrics. Chem. Mater. **33**(17), 7008–7016 (2021) https://doi.org/10.1021/acs.chemmater.1c02061

[31] Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A.: Automatic chemical design using a data-driven

continuous representation of molecules. ACS Cent. Sci. **4**(2), 268–276 (2018) https://doi.org/10.1021/acscentsci.7b00572 . PMID: 29532027

[32] Munson, B.P., Chen, M., Bogosian, A., Kreisberg, J.F., Licon, K., Abagyan, R., Kuenzi, B.M., Ideker, T.: De novo generation of multi-target compounds using deep generative chemistry. Nat. Commun. **15**(1), 3636 (2024) https://doi.org/10.1038/s41467-024-47120-y

[33] Xu, M., Powers, A.S., Dror, R.O., Ermon, S., Leskovec, J.: Geometric latent diffusion models for 3D molecule generation. In Proceedings of the 40th International Conference on Machine Learning (ICML) **202**, 38592–38610 (2023)

[34] Weiss, T., Mayo Yanes, E., Chakraborty, S., Cosmo, L., Bronstein, A.M., Gershoni-Poranne, R.: Guided diffusion for inverse molecular design. Nat. Comput. Sci. **3**(10), 873–882 (2023) https://doi.org/10.1038/s43588-023-00532-0

[35] Bagal, V., Aggarwal, R., Vinod, P.K., Priyakumar, U.D.: Molgpt: Molecular generation using a transformer-decoder model. J. Chem. Inf. Model. **62**(9), 2064–2076 (2022) https://doi.org/10.1021/acs.jcim.1c00600

[36] Chang, J., Ye, J.C.: Bidirectional generation of structure and properties through a single molecular foundation model. Nat. Commun. **15**(1), 2323 (2024) https://doi.org/10.1038/s41467-024-46440-3

[37] Swanson, K., Liu, G., Catacutan, D.B., Arnold, A., Zou, J., Stokes, J.M.: Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. Nat. Mach. Intell. **6**(3), 338–353 (2024) https://doi.org/10.1038/s42256-024-00809-7

[38] Ma, R., Zhang, H., Luo, T.: Exploring High Thermal Conductivity Amorphous Polymers Using Reinforcement Learning. ACS Appl. Mater. Interfaces **14**(13), 15587–15598 (2022) https://doi.org/10.1021/acsami.1c23610

[39] Liu, D.-F., Zhang, Y.-X., Dong, W.-Z., Feng, Q.-K., Zhong, S.-L., Dang, Z.-M.: High-temperature polymer dielectrics designed using an invertible molecular graph generative model. J. Chem. Inf. Model. **63**(24), 7669–7675 (2023) https://doi.org/10.1021/acs.jcim.3c01572

[40] Ahmad, W., Simon, E., Chithrananda, S., Grand, G., Ramsundar, B.: Chemberta-2: Towards chemical foundation models. arXiv preprint arXiv:2209.01712 (2022)

[41] 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 **57**(8), 3515–3528 (2024) https://doi.org/10.1021/acs.macromol.4c00508

[42] Ma, R., Luo, T.: PI1M: A Benchmark Database for Polymer Informatics. J. Chem. Inf. Model. **60**(10), 4684–4690 (2020) https://doi.org/10.1021/acs.jcim.0c00726

[43] Webb, M.A., Jackson, N.E., Gil, P.S., Pablo, J.J.: Targeted sequence design within the coarse-grained polymer genome. Sci. Adv. **6**(43), 6216 (2020) https://doi.org/10.1126/sciadv.abc6216

[44] Tao, L., He, J., Munyaneza, N.E., Varshney, V., Chen, W., Liu, G., Li, Y.: Discovery of multi-functional polyimides through high-throughput screening using explainable machine learning. Chem. Eng. J. **465**, 142949 (2023) https://doi.org/10.1016/j.cej.2023.142949

[45] 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, 10–1039209272 (2023) https://doi.org/10.1039/D2TA09272G

[46] 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. **5**(9), 1523–1531 (2019) https://doi.org/10.1021/acscentsci.9b00476 https://doi.org/10.1021/acscentsci.9b00476

[47] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 1–67 (2020)

[48] Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminformatics **7**, 20 (2015)

[49] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J. Cheminformatics **1**(1), 8 (2009)