

polyBART: A Chemical Linguist for Polymer Property Prediction and Generative Design

Anagha Savit¹, Harikrishna Sahu¹, Shivank Shukla¹,
Wei Xiong¹, Rampi Ramprasad¹

¹ School of Materials Science and Engineering, Georgia Institute of Technology

Correspondence: rampi.ramprasad@mse.gatech.edu

Abstract

Designing polymers for targeted applications and accurately predicting their properties is a key challenge in materials science owing to the vast and complex polymer chemical space. While molecular language models have proven effective in solving analogous problems for molecular discovery, similar advancements for polymers are limited. To address this gap, we propose polyBART, a language model-driven polymer discovery capability that enables rapid and accurate exploration of the polymer design space. Central to our approach is Pseudo-polymer SELFIES (PSELFIES), a novel representation that allows for the transfer of molecular language models to the polymer space. polyBART is, to the best of our knowledge, the first language model capable of bidirectional translation between polymer structures and properties, achieving state-of-the-art results in property prediction and design of novel polymers for electrostatic energy storage. Further, polyBART is validated through a combination of both computational and laboratory experiments. We report what we believe is the first successful synthesis and validation of a polymer designed by a language model, predicted to exhibit high thermal degradation temperature and confirmed by our laboratory measurements. Our work presents a generalizable strategy for adapting molecular language models to the polymer space and introduces a polymer foundation model, advancing generative polymer design that may be adapted for a variety of applications.

1 Introduction

Polymers play a crucial role in our daily lives, serving as essential constituents of countless materials and products that we rely on. Designing novel application-specific polymers, however, continues to pose a significant challenge owing to the vast polymer chemical space. The application of Machine Learning (ML) in this space (Jørgensen et al.,

2018; Batra et al., 2021; Tran et al., 2024) has made significant strides in addressing the forward problem of predicting material properties from polymer structures (Doan Tran et al., 2020; Andraju et al., 2022). In contrast, progress on the inverse problem, namely, rapidly designing polymers that meet target property requirements, has been far more limited, largely due to the inherent difficulty of generating chemically valid polymer structures with traditional ML.

Language models, owing to their generative capabilities, have emerged as an effective strategy for addressing both forward and inverse problems in the molecular space. Drawing inspiration from foundation models in Natural Language Processing (NLP), molecular foundation models (Edwards et al., 2022; Bagal et al., 2022) are pretrained on the 'language' of molecules, typically represented as SMILES (Weininger, 1988), SMARTS (Daylight Chemical Information Systems), or SELFIES (Krenn et al., 2020) strings. These models are subsequently fine-tuned on downstream tasks, such as property prediction, hence solving the forward problem. The problem of generative design is often achieved by navigating the latent space learned during pretraining. The achievements of language models for molecules highlight a key opportunity for extending similar methodologies to polymers. Particularly, the structural and representational parallels between the molecular and polymer domains present a unique possibility of adapting existing molecular foundation models to the polymer space.

Recognizing this, we develop polyBART by strategically leveraging the chemical priors learned by existing molecular foundation models. To accomplish this, we introduce a novel representation, Pseudo-polymer SELFIES (PSELFIES), an extension of SELFIES (Krenn et al., 2020) adapted for polymers. SELFIES (Self-Referencing Embedded Strings) is a robust molecular string representation that guarantees 100% syntactic validity, mean-

ing every SELFIES string corresponds to a valid molecule. From a chemical standpoint, PSELFIES allow us to represent polymers in a format that mirrors molecular syntax, enabling direct compatibility with existing molecular language models and facilitating seamless continued pretraining. Further details on this representation is found in Section 3.1.

We develop polyBART through the continued pretraining of SELFIES-TED (IBM Research), an encoder-decoder model based on BART (Lewis et al., 2019) and designed for molecular representations. PolyBART is a unique unifying model capable of solving both forward and inverse problems in polymer informatics. Our work makes the following key contributions:

- We introduce PSELFIES and develop polyBART, a polymer foundation model that is, to our knowledge, the first application of language models to generative polymer design.
- We provide comprehensive computational validation of polyBART, demonstrating strong performance in both polymer property prediction and generation.
- We report laboratory synthesis and testing of a polymer predicted by polyBART, which, to our knowledge, is the first successfully validated polymer design guided by a language model. Notably, upon thermal property testing, we find that the experimental measurements align with the model’s predictions.

2 Related Work

2.1 Chemical Foundation Models

Recent work has successfully adapted transformer architectures (Vaswani et al., 2017) to the molecular domain, achieving impressive results (Ross et al., 2022). Early work has focused on adapting well-established transformer backbones, originally developed for natural language understanding, to molecule tasks by training them on molecular string representations. BERT (Devlin et al., 2019) has been widely used for this purpose, forming the foundation for models like SMILES-BERT (Wang et al., 2019) and Mol-BERT (Li and Jiang, 2021). Similarly, ChemBERTa (Chithrananda et al., 2020) builds on the RoBERTa architecture (Liu et al., 2019) and is pretrained on millions of SMILES strings. Later works have explored the application of Large Language Models (LLMs) to these

tasks. Unlike the earlier transformer-based methods, LLMs benefit from their exposure to vast amounts of data, eliminating the need for the pre-training step. For instance, Gruver et al. (2024) demonstrated that instruction fine-tuned LLaMA models (Touvron et al., 2023) can effectively generate crystal structures. Similarly, Jablonka et al. (2024) showed that GPT-3 (Brown et al., 2020) can be fine-tuned using natural language to perform a wide range of tasks in chemistry and materials science.

Alongside these developments, other works have shown that transformer models can also be used to generate molecular structures and not just to predict their properties. Models like MolT5 (Edwards et al., 2022) and MolGPT (Bagal et al., 2022) enable de novo molecule generation, MolT5 from textual descriptions, and MolGPT through next-token prediction. The SELFIES-TED model generates molecules by exploring its learned latent space. Diffusion methods have also emerged as a strong candidate for molecular generation, as demonstrated by GenMol (Lee et al., 2025) and TGM-DLM (Gong et al., 2024). Our polymer foundation model builds upon SELFIES-TED and hence adopts a parallel strategy for property prediction and generation.

2.2 Polymer Informatics via NLP

ML algorithms have been applied in the polymer space for quite some time (Chen et al., 2021). Similar to molecules, a common approach in polymer informatics is to represent polymers numerically and use ML to connect these representations to points in the property space. A range of approaches, including regression algorithms (Doan Tran et al., 2020; Tao et al., 2021) and neural networks (Kuenneth et al., 2021b; Kuenneth et al., 2022) have been explored for this purpose. Graph neural networks (GNNs) are also widely used in polymer modeling due to their ability to capture both local and global topological features of polymer graphs. Gur-nani et al. (2023) developed polyGNN, a multitask polymer GNN model for polymer property prediction, demonstrating strong performance across numerous properties. While the application of NLP in this space is relatively nascent, it has already seen significant success. In property prediction tasks, language models are typically used to generate numerical embeddings of polymer representations, which are then mapped to properties using approaches similar to traditional polymer informat-

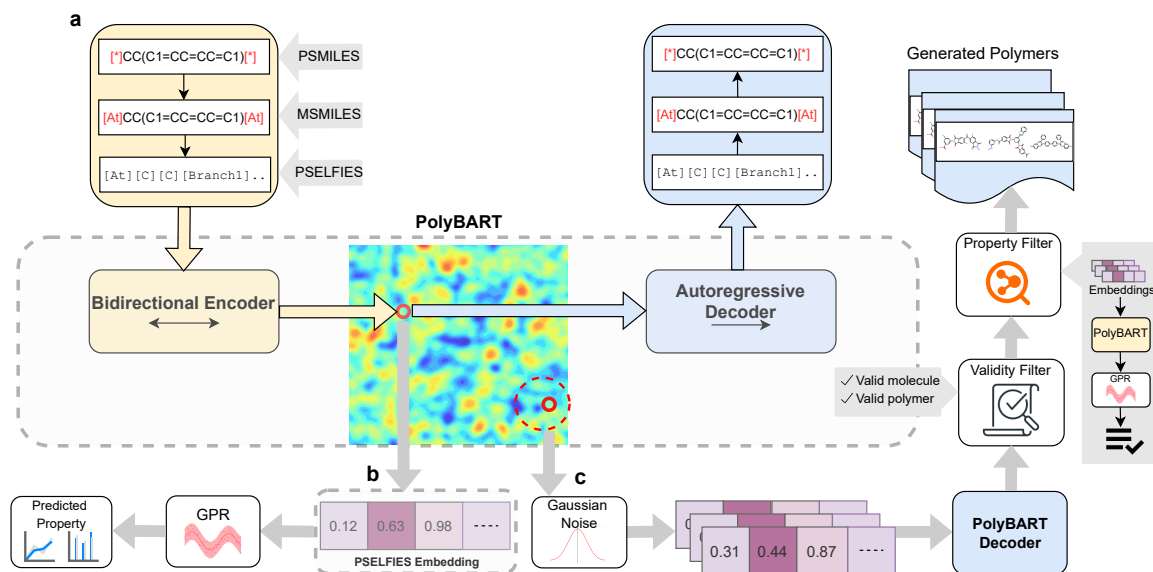


Figure 1: Overview of the polyBART pipeline. (a) The encoder-decoder model is pretrained using a Masked Language Modeling (MLM) objective to learn the language of PSELFIES and construct the latent space. (b) For property prediction, the trained encoder generates embeddings for polymers with known properties, which are then mapped to target values using Gaussian Process Regressor (GPR). (c) For generating new structures, Gaussian noise is added to the learned embeddings, and the decoder generates candidate polymers, which are subsequently filtered based on property and synthesizability criteria.

ics pipelines (Agarwal et al., 2025). Kuenneth and Ramprasad (2023) developed polyBERT by pre-training deBERTA (He et al., 2021) on millions of SMILES strings. Similar polymer language models have also been developed using alternative architectures, such as RoBERTa in TransPolymer (Xu et al., 2023) and T5 (Raffel et al., 2023) in PolyNC (Qiu et al., 2024).

For the inverse problem of generating polymer structures, early efforts have leveraged Variational Autoencoders (VAEs) (Batra et al., 2020). However, the lack of explicit chemical syntax in VAE architectures leads to invalid generations. Transformers can learn context-aware representations of molecular structures (Dollar et al., 2021), with each attention head learning to focus on distinct substructures. By leveraging SELFIES (or PSELFIES in the present case), which guarantee validity, and language models that can capture structured sequences and syntactic patterns, we achieve state-of-the-art results in the generation of novel, valid, and property-aligned polymers.

3 polyBART: A Polymer Foundation Model

In this section we detail the PSELFIES representation and introduce polyBART. The overall model

pipeline is shown in Figure 1.

3.1 PSELFIES

To ensure compatibility with molecular language models, we introduce PSELFIES, a novel representation of polymers. PSELFIES builds upon several representations used in cheminformatics. SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988) is a widely adopted notation for encoding molecules as strings. Polymer SMILES (PSMILES) extends SMILES to polymers by representing the two terminal endpoints of homopolymer repeat units with asterisks ([*]) to indicate open-chain ends. However, [*] is commonly used as a placeholder atom in molecular representations rather than specifically denoting polymer termini, and it is not yet incorporated into the SELFIES representation.

To address this limitation, we first convert PSMILES into Molecule SMILES (MSMILES), replacing [*] with actual atomic representations. The conversion begins by identifying the terminal atoms and bond types associated with [*], then joining the polymer ends to form a cyclic structure, followed by eliminating the [*]s. To ensure consistency and to remove biases arising from multiple valid PSMILES representations, we canonicalize the cyclic structure. Next, we strategically cleave a

Model	T_g (K)	T_d (K)	T_m (K)	E_{gc} (eV)	E_{gb} (eV)	E_{ea} (eV)	E_i (eV)
SELFIES-TED _{large}	62.58 ± 1.94	96.98 ± 2.14	73.77 ± 2.97	0.95 ± 0.04	0.78 ± 0.09	0.37 ± 0.06	0.50 ± 0.06
polyBERT	38.41 ± 1.92	67.45 ± 2.06	56.25 ± 3.09	0.55 ± 0.02	<u>0.72 ± 0.06</u>	0.35 ± 0.05	0.49 ± 0.04
polyBART _{small}	<u>39.92 ± 1.27</u>	71.56 ± 1.64	<u>57.71 ± 2.36</u>	<u>0.60 ± 0.02</u>	0.68 ± 0.08	0.35 ± 0.05	0.51 ± 0.06
polyBART _{large}	40.32 ± 1.16	<u>71.44 ± 1.93</u>	58.02 ± 2.36	0.61 ± 0.02	0.73 ± 0.07	0.33 ± 0.04	0.49 ± 0.08
BART _{small}	43.35 ± 1.25	73.35 ± 1.94	62.03 ± 2.83	-	-	-	-
LLaMA-3-8B	58.13	95.28	71.91	-	-	-	-

Table 1: Performance on the property prediction tasks. The best results are highlighted in bold and second-best are underlined.

tokens in the input sequence randomly masked. Detailed descriptions of the model architecture and training setup can be found in Appendix C. Ultimately, our approach allows us to leverage the extensive chemical knowledge encoded in the base models, pretrained on billions of molecules, while progressively specializing the models to learn the representations of polymers. Through further training on millions of PSELFIES strings, polyBART thus obtains a deep understanding of the grammar and syntax that govern the polymer chemical language.

4 Computational Experiments

In this section, we present results demonstrating the effectiveness of polyBART on: (i) property prediction, (ii) hypothetical polymer generation, and (iii) property-guided never-seen-before polymer generation. We also highlight the ability of PSELFIES in leveraging pretrained knowledge through a comparative study.

4.1 Property Prediction

Task: Polymer property prediction is formulated as a supervised learning task in our pipeline and accomplished as follows: first, all the PSMILES strings from the property datasets are converted to their PSELFIES representations. These are then passed through the polyBART encoder to generate numerical embeddings. To reduce dimensionality, we apply Principal Component Analysis (PCA) to the resulting PSELFIES embeddings. The reduced representations are then used to train a supervised learning model, specifically, Gaussian Process Regression (GPR) with a Radial Basis Function (RBF) kernel. We use a 5-fold Cross-Validation (CV) strategy to assess the performance and generalizability of our GPR models.

Datasets: We evaluate polyBART on critical

thermal properties of polymers, namely, the glass transition temperature (T_g), thermal degradation temperature (T_d), and melting temperature (T_m) (Kuenneth et al., 2021a). Thermal property data is collected from the PolyInfo repository (Otsuka et al., 2011) and is derived from experimental measurements reported in the existing literature. We also assess performance on electronic properties, including linear chain bandgap (E_{gc}), bulk polymer bandgap (E_{gb}), electron affinity (E_{ea}), and ionization energy (E_i) values computed using Density Functional Theory (DFT). Additional information on the datasets is detailed in Appendix D.

Baselines: We compare the performance of polyBART to the baseline SELFIES-TED model. Due to the absence of the [At] token in SELFIES-TED_{small}, our comparison is restricted to SELFIES-TED_{large}. In addition, we benchmark polyBART with polyBERT, current state-of-the-art transformer-based model for polymer property prediction. It is important to note that, given the novel bidirectional capabilities of polyBART, we are limited in comparison to unidirectional models developed for property prediction. To ensure a fair comparison of embedding quality, we use GPR across all models. We additionally compare polyBART with LLaMA, with details regarding the fine-tuning experiment given in Appendix E.

Results: Table 1 presents the performance of polyBART_{small} and polyBART_{large}, reporting the average RMSE across five CV test sets along with the corresponding standard deviations. We observe that our bidirectional polyBART performs on par with, and in many cases surpasses, polyBERT. Notably, while matching the predictive accuracy of the unidirectional polyBERT, the bidirectional polyBART offers the added advantage of generating entirely new polymer candidates not present in the training data. Additionally, polyBART_{small} and

polyBART_{large} outperform the baseline SELFIES-TED_{large}, achieving average RMSE reductions of 19.4% and 19.7%, respectively. These consistent performance gains underscore the ability of PSELFIES to specialize molecular language models to the polymer domain. We also notice that polyBART outperforms LLaMA-3 on thermal property prediction. Overall, our results confirm the capability of polyBART as a powerful polymer property predictor.

4.2 Impact of Transfer Learning

To further demonstrate the importance of PSELFIES in adapting molecular foundation models to the polymer space, we perform a comparative study. Specifically, we develop a model, BART_{small}, which shares the same architecture, training data, and configuration as polyBART_{small}, but is trained from scratch rather than initialized from SELFIES-TED. This setup allows us to isolate the effect of transfer learning using PSELFIES.

We compare the performance of BART_{small} against polyBART_{small} on property prediction tasks across the three thermal properties, with results summarized in Table 1. We observe that polyBART_{small} consistently outperforms BART_{small} in all tasks. These performance gains highlight the critical role of our PSELFIES framework. By leveraging pretrained knowledge from SELFIES-TED, polyBART benefits from a strong initialization, leading to significant improvements over training from scratch. Overall, we confirm that rather than starting from random weights, initializing with a pretrained molecular model via PSELFIES provides an effective foundation for developing polymer language models.

4.3 Hypothetical Polymer Generation

polyBART’s generative capability, enabled by its decoder, is a key innovation of our work, setting it apart from existing polymer language models (Agarwal et al., 2025), which have been limited to unidirectional property prediction. To generate novel polymer structures, the PSELFIES sequence is first passed through the encoder to obtain a latent embedding. To explore the latent space, we apply Gaussian noise to this embedding using an n-fold sampling strategy: noise is added n times to generate n noised embeddings. These noisy embeddings are then decoded one by one using the polyBART decoder, and if a valid polymer is generated within

these samples, it is retained. A grid search is used to tune the noise level for novelty and validity, details of which can be found in Appendix F.

We evaluate the generative capabilities of polyBART using a test set of 10,000 PSELFIES, all of which are unseen during the pretraining phase. We begin by examining the validity of the generated polymer structures. Specifically, we introduce two measures of validity: (i) molecule validity and (ii) polymer validity. Molecule validity checks whether the generated PSELFIES correspond to a chemically valid molecule, and is determined using RDKit. Polymer validity assesses whether the generated PSELFIES represent a structurally valid polymer by examining the placement and single-bond character of the [At] tokens. Since the [At] tokens denote the endpoints of the polymer repeat unit, a valid polymer must contain exactly two such tokens, each with a valency of one. The polymer validity metric enforces this requirement. Next, we assess the novelty of the valid generated structures, where a structure is considered novel if it passes both validity tests and is different from all input examples. Finally, we assess the quality of the generated structures using Fréchet ChemNet Distance (FCD) (Preuer et al., 2018) and Internal Diversity (IntDiv_p) metrics (Benhenda, 2017). FCD quantifies the similarity between the distributions of the generated and input structures based on the Euclidian distance between their respective embeddings. A higher FCD score indicates a greater divergence between the two distributions, suggesting that the generated structures differ from the input set. IntDiv_p evaluates the internal diversity of the generated structures, capturing the model’s tendency to produce structurally distinct outputs. We report both IntDiv₁ (p = 1) and IntDiv₂ (p = 2) scores, where higher values of these scores reflect greater diversity in the generated structures. Although these metrics were originally designed for molecules, by representing polymers as pseudomolecules using PSELFIES, we can extend them to our case. See Appendix G for a detailed description of the metrics.

Results for the generation task are provided in Table 2. We find that polyBART_{small} and polyBART_{large} generate valid molecules in 100% of cases, and valid polymers in 91% and 98% of cases, respectively, demonstrating that the models have effectively learned the underlying grammar and structural rules in writing polymers. High novelty scores demonstrate polyBART’s ability to ex-

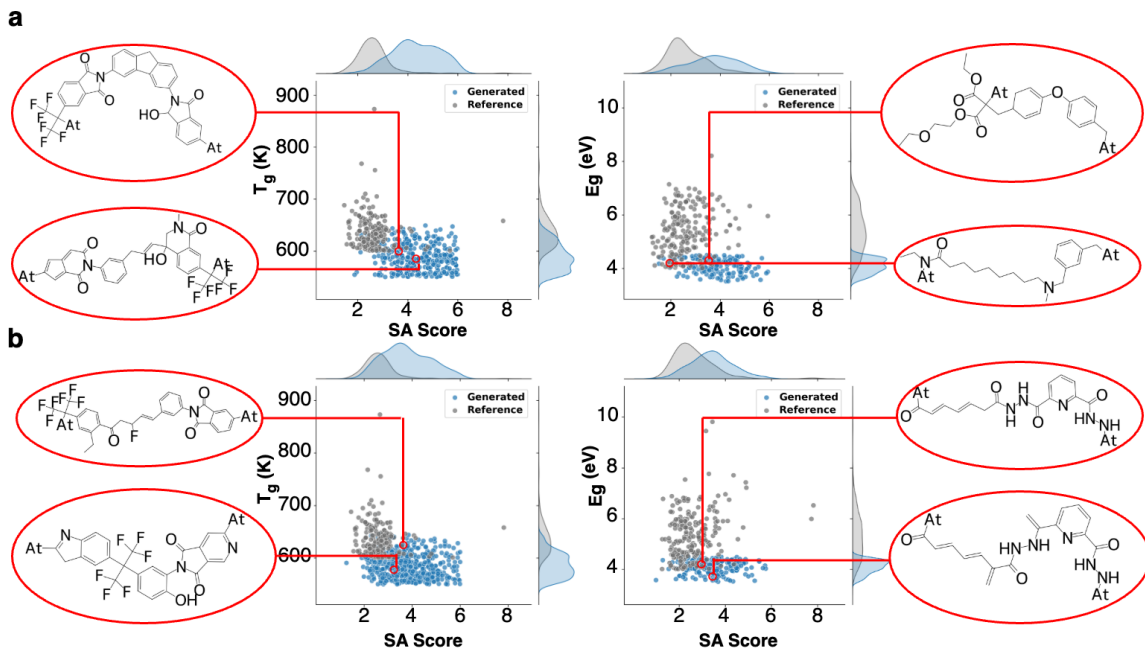


Figure 3: Representative chemical structures generated by (a) polyBART_{small} and (b) polyBART_{large}, selected for their high T_g , high E_g , and ease of synthesizability. The selected examples highlight polyBART’s ability to generate thermally and electronically robust polymers.

Model	Molecule Validity	Polymer Validity	Novelty	FCD	IntDiv ₁	IntDiv ₂
polyBART _{small}	1.000	0.913	0.867	3.968	0.878	0.875
polyBART _{large}	1.000	0.985	0.793	1.833	0.872	0.868

Table 2: Generative performance of polyBART_{small} and polyBART_{large} across key evaluation metrics.

plore the latent space and generate thousands of never-seen-before structures. The FCD and IntDiv_p scores of the generated distributions confirm the models’ capabilities to generate a broad and diverse range of polymer structures.

4.4 Property-conditioned Polymer Generation

Beyond general structure generation, a key capability of polyBART is its ability to generate novel polymers conditioned on desired target properties. At a high level, our approach enables property-guided generation by mapping the latent space of polymer embeddings to the property space and exploring regions with desired property values. We begin by identifying polymers that fall within a specified target range for a given property. These polymers are then passed through polyBART’s encoder to obtain embeddings. As in the general generation setting, we apply an n-fold Gaussian noise sampling strategy: each embedding is perturbed n times to generate n candidate embeddings, which are subsequently decoded and checked for validity.

In this way, we are able to effectively explore neighborhoods in the latent space with desirable property values. The resulting candidate polymers are subjected to two filtering steps: (i) property filtering and (ii) Synthetic Accessibility (SA) score (Ertl and Schuffenhauer, 2009) filtering. To filter based on property, the generated candidates are passed through the GPR models developed for property prediction, and only those within a threshold range of the target property are retained. To assess synthesizability, we compute the SA score, a measure ranging from 1 (easy to make) to 10 (very difficult to make). For this, PSELFIES strings are converted to SMILES, with the At atom replaced by hydrogen (H). The SA score is then calculated, and candidates with a score of 6 or lower are retained as viable candidates. Additional details are provided in Appendix H.

We use polyBART to generate novel polymer structures under two extreme conditions: (i) high temperatures and (ii) high electric fields. We make the assumption that these behaviors correlate with

Metric	T_g		E_{gc}	
	polyBART _{small}	polyBART _{large}	polyBART _{small}	polyBART _{large}
# Valid Generations	743	1131	739	1514
# Novel	715	933	558	660
# Property Filtered	396	543	168	190
# SA Score Filtered	353	511	160	189

Table 3: Performance of polyBART_{small} and polyBART_{large} on T_g and E_{gc} guided generation of novel polymer candidates, based on 183 and 200 reference polymers, respectively.

two critical polymer properties, specifically, T_g and E_{gc} . Accordingly, we define our target property thresholds as (i) $T_g > 600$ K and (ii) $E_{gc} > 4$ eV. High T_g polymers are sought for their thermal integrity at elevated temperatures, while polymers with high E_{gc} values provide broad electrical stability with minimal dielectric loss, making them promising candidates for high-energy density capacitor dielectric applications (Tan et al., 2014; Zhou et al., 2018; Gurnani et al., 2024).

As shown in Table 3, polyBART_{small} and polyBART_{large} are able to successfully generate hundreds of novel polymers that not only meet the desired property criteria but are also predicted to be synthetically accessible. Examples of novel structures generated by our models are illustrated in Figure 3. Additionally, visualizations of the property-conditioned latent space can be found in Appendix I.

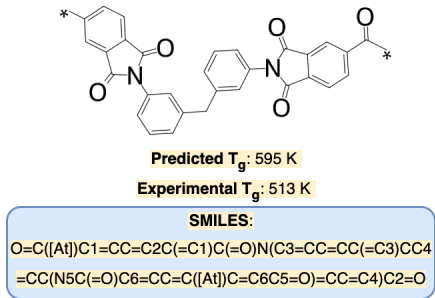


Figure 4: A high T_g polymer structure generated by polyBART and synthesized in the lab.

5 Validation of Design via Laboratory Synthesis and Testing

Given the real-world implications of our work, it is important to extend our evaluation of polyBART beyond computational benchmarks. Thus, we perform experimental synthesis and characterization of a polymer candidate predicted by polyBART_{small}. We select one polymer with a high predicted T_g and

a favorable SA score and successfully synthesize it in the lab. Following synthesis, we perform T_g measurements to evaluate its thermal properties. The experimental protocol is detailed in Appendix J. Figure 4 presents the chemical structure of the synthesized polymer, along with the predicted and experimentally measured T_g values. We confirm that the experimentally observed T_g aligns well with the prediction. This successful synthesis of a high- T_g polymer firmly establishes polyBART’s ability to guide real-world polymer design.

6 Conclusion

In this work, we propose PSELFIES, a novel representation of polymers that enables the transfer of learned knowledge from molecular language models to the polymer space. We develop polyBART_{small} and polyBART_{large}, two polymer foundation models created through the continued pretraining of SELFIES-TED_{small} and SELFIES-TED_{large}, respectively. Our results show that polyBART performs comparably to state-of-the-art approaches in property prediction. Moreover, polyBART uniquely combines property prediction with generative design in a single unified framework and consistently produces numerous novel polymer structures. Generated candidates not only satisfy target property criteria but also exhibit synthetic accessibility. By successfully synthesizing one such predicted polymer and experimentally confirming our predictions, we illustrate that language models can reliably drive real-world advances in polymer design. Our work enables true bidirectional translation between polymer structures and properties and highlights the potential of language models in advancing polymer informatics.

7 Limitations

While PSELFIES successfully enables the adaptation of the SELFIES-TED molecular language

model to the polymer space, its generalizability to other molecular foundation models, based on different architectures and pretraining strategies remains yet to be seen. Future work is needed to evaluate how well PSELFIES integrates with alternative models and whether similar performance gains can be achieved across methods.

Additionally, while polyBART demonstrates strong performance on thermal and electronic properties, its applicability to a broader range of polymer properties and simultaneous optimization of multiple properties must be explored to fully determine its predictive and generative power. Finally, polyBART is currently limited to homopolymers and does not yet support the representation or generation of copolymers, polymer blends, or polymer composites/formulations. These aspects will be the subject of future inquiry.

8 Acknowledgements

The authors thank the Office of Naval Research (ONR) for financial sponsorship of this work through Grant N00014-23-1-2279. The authors are also grateful for the usage of the Gutekunst lab for synthesis and characterization of one of the designed polymers.

References

- Sakshi Agarwal, Akhlak Mahmood, and Rampi Ramprasad. 2025. [Polymer solubility prediction using large language models](#). *ACS Materials Letters*, pages 2017–2023.
- Nagababu Andraju, Greg W. Curtzwiler, Yun Ji, Evguenii Kozliak, and Prakash Ranganathan. 2022. [Machine-learning-based predictions of polymer and postconsumer recycled polymer properties: A comprehensive review](#). *ACS Applied Materials & Interfaces*, 14(38):42771–42790. PMID: 36102317.
- Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. [Molgpt: Molecular generation using a transformer-decoder model](#). *Journal of Chemical Information and Modeling*, 62(9):2064–2076. PMID: 34694798.
- Rohit Batra, Hanjun Dai, Tran Doan Huan, Lihua Chen, Chiho Kim, Will R. Gutekunst, Le Song, and Rampi Ramprasad. 2020. [Polymers for extreme conditions designed using syntax-directed variational autoencoders](#). *Chemistry of Materials*, 32(24):10489–10500.
- Rohit Batra, Le Song, and Rampi Ramprasad. 2021. [Emerging materials intelligence ecosystems propelled by machine learning](#). *Nature Reviews Materials*, 6(8):655–678.
- Mostapha Benhenda. 2017. [Chemgan challenge for drug discovery: can ai reproduce natural chemical diversity?](#) *Preprint*, arXiv:1708.08227.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Lihua Chen, Ghanshyam Pilania, Rohit Batra, Tran Doan Huan, Chiho Kim, Christopher Kuenneth, and Rampi Ramprasad. 2021. [Polymer informatics: Current status and critical next steps](#). *Materials Science and Engineering: R: Reports*, 144:100595.
- Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. [Chemberta: Large-scale self-supervised pretraining for molecular property prediction](#). *Preprint*, arXiv:2010.09885.
- Daylight Chemical Information Systems. SMARTS — A Language for Describing Molecular Patterns. <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- Huan Doan Tran, Chiho Kim, Lihua Chen, Anand Chandrasekaran, Rohit Batra, Shruti Venkatram, Deepak Kamal, Jordan P. Lightstone, Rishi Gurnani, Pranav Shetty, Manav Ramprasad, Julia Laws, Madeline Shelton, and Rampi Ramprasad. 2020. [Machine-learning predictions of polymer properties with polymer genome](#). *Journal of Applied Physics*, 128(17):171104.
- Orion Dollar, Nisarg Joshi, David A. C. Beck, and Jim Pfandtner. 2021. [Attention-based generative models for de novo molecular design](#). *Chem. Sci.*, 12:8362–8372.
- Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. [Translation between molecules and natural language](#). *Preprint*, arXiv:2204.11817.
- Peter Ertl and Ansgar Schuffenhauer. 2009. [Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions](#). *Journal of Cheminformatics*, 1:8.
- Zhishuai Geng, Jaeman J. Shin, Yumeng Xi, and Craig J. Hawker. 2021. [Click chemistry strategies for the accelerated synthesis of functional macromolecules](#). *Journal of Polymer Science*, 59(11):963–1042.

- Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024. [Text-guided molecule generation with diffusion language model](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):109–117.
- Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ulissi. 2024. [Fine-tuned language models generate stable inorganic materials as text](#). *Preprint*, arXiv:2402.04379.
- Rishi Gurnani, Christopher Kuenneth, Aubrey Toland, and Rampi Ramprasad. 2023. [Polymer informatics at scale with multitask graph neural networks](#). *Chemistry of Materials*, 35(4):1560–1567.
- Rishi Gurnani, Stuti Shukla, Deepak Kamal, Chao Wu, Jing Hao, Christopher Kuenneth, Pritish Akuljkar, Ashish Khomane, Robert Daniels, Ajinkya A. Deshmukh, Yang Cao, Gregory Sotzing, and Rampi Ramprasad. 2024. [Ai-assisted discovery of high-temperature dielectrics for energy storage](#). *Nature Communications*, 15(1):6107.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). *Preprint*, arXiv:2006.03654.
- IBM Research. selfies-ted. <https://huggingface.co/ibm-research/materials.selfies-ted>.
- Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. [Leveraging large language models for predictive chemistry](#). *Nature Machine Intelligence*, 6(2):161–169.
- Peter Bjørn Jørgensen, Murat Mesta, Suranjan Shil, Juan María García Lastra, Karsten Wedel Jacobsen, Kristian Sommer Thygesen, and Mikkel N. Schmidt. 2018. [Machine learning-based screening of complex molecules for polymer solar cells](#). *The Journal of Chemical Physics*, 148(24):241735.
- Seonghwan Kim, Charles M. Schroeder, and Nicholas E. Jackson. 2023. [Open macromolecular genome: Generative design of synthetically accessible polymers](#). *ACS Polymers Au*, 3(4):318–330.
- Hartmuth C. Kolb, M. G. Finn, and K. Barry Sharpless. 2001. [Click chemistry: Diverse chemical function from a few good reactions](#). *Angewandte Chemie International Edition*, 40(11):2004–2021.
- Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. [Self-referencing embedded strings \(selfies\): A 100% robust molecular string representation](#). *Machine Learning: Science and Technology*, 1(4):045024.
- Christopher Kuenneth, Jessica Lalonde, Babetta L. Marone, Carl N. Iverson, Rampi Ramprasad, and Ghan-shyam Pilania. 2022. [Bioplastic design using multi-task deep neural networks](#). *Communications Materials*, 3(1):96.
- Christopher Kuenneth, Arunkumar Chitteth Rajan, Huan Tran, Lihua Chen, Chiho Kim, and Rampi Ramprasad. 2021a. [Polymer informatics with multi-task learning](#). *Patterns*, 2(4):100238.
- Christopher Kuenneth and Rampi Ramprasad. 2023. [polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics](#). *Nature Communications*, 14(1):4099.
- Christopher Kuenneth, William Schertzer, and Rampi Ramprasad. 2021b. [Copolymer informatics with multitask deep neural networks](#). *Macromolecules*, 54(13):5957–5961.
- Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saeed Paliwal, Weili Nie, and Arash Vahdat. 2025. [Genmol: A drug discovery generalist with discrete diffusion](#). *Preprint*, arXiv:2501.06158.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *Preprint*, arXiv:1910.13461.
- Juncai Li and Xiaofei Jiang. 2021. [Mol-bert: An effective molecular representation with bert for molecular property prediction](#). *Wireless Communications and Mobile Computing*, 2021(1):7181815.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Mitsuru Ohno, Yoshihiro Hayashi, Qi Zhang, Yu Kaneko, and Ryo Yoshida. 2023. [Smipoly: Generation of a synthesizable polymer virtual library using rule-based polymerization reactions](#). *Journal of Chemical Information and Modeling*, 63(17):5539–5548. PMID: 37604495.
- Shingo Otsuka, Isao Kuwajima, Junko Hosoya, Yibin Xu, and Masayoshi Yamazaki. 2011. [Polyinfo: Polymer database for polymeric materials design](#). In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, pages 22–29.
- Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. [Fréchet chemnet distance: A metric for generative models for molecules in drug discovery](#). *Preprint*, arXiv:1803.09518.
- Haoke Qiu, Lunyang Liu, Xuepeng Qiu, Xuemin Dai, Xiangling Ji, and Zhao-Yan Sun. 2024. [Polync: a natural and chemical language model for the prediction of unified polymer properties](#). *Chem. Sci.*, 15:534–544.

- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Preprint*, arXiv:1910.10683.
- Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. [Large-scale chemical language representations capture molecular structure and properties](#). *Nature Machine Intelligence*, 4(12):1256–1264.
- Daniel Tan, Lili Zhang, Qin Chen, and Patricia Irwin. 2014. [High-temperature capacitor polymer films](#). *Journal of Electronic Materials*, 43(12):4569–4575.
- Lei Tao, Vikas Varshney, and Ying Li. 2021. [Benchmarking machine learning models for polymer informatics: An example of glass transition temperature](#). *Journal of Chemical Information and Modeling*, 61(11):5395–5413. PMID: 34662106.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Huan Tran, Rishi Gurnani, Chiho Kim, Ghanshyam Pillania, Ha-Kyung Kwon, Ryan P. Lively, and Rampi Ramprasad. 2024. [Design of functional and sustainable polymers assisted by artificial intelligence](#). *Nature Reviews Materials*, 9(12):866–886.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. [Smiles-bert: Large scale unsupervised pre-training for molecular property prediction](#). In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '19*, page 429–436, New York, NY, USA. Association for Computing Machinery.
- David Weininger. 1988. [Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules](#). *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.
- Changwen Xu, Yuyang Wang, and Amir Barati Farmani. 2023. [Transpolymer: a transformer-based language model for polymer property predictions](#). *npj Computational Materials*, 9(1):64.
- Robert J. Young and Peter A. Lovell. 2011. Chapters 3 and 7. In *Introduction to Polymers*, 3 edition, pages 21–60, 151–189. CRC Press.
- Yao Zhou, Qi Li, Bin Dang, Yang Yang, Tao Shao, He Li, Jun Hu, Rong Zeng, Jinliang He, and Qing Wang. 2018. [A scalable, high-throughput, and environmentally benign approach to polymer dielectrics exhibiting significantly improved capacitive performance at high temperatures](#). *Advanced Materials*, 30(49):1805672.

Appendix

A Representations

Representation	Description
SMILES	Simplified Molecular Input Line Entry System is a popular notation for encoding molecules as strings.
PSMILES	Polymer SMILES is an extension of SMILES for polymers that uses asterisks ([*]) to indicate endpoints of homopolymer repeat units.
MSMILES	Molecule SMILES is a pseudo-molecular SMILES derived from PSMILES by ring closure, canonicalization, and bond cleavage.
SELFIES	Self-Referencing Embedded Strings is a robust representation of molecules that guarantees the validity of every generation.
PSELFIES	Polymer SELFIES is a novel polymer-specific extension of SELFIES that is introduced in this work and used to develop polyBART.

Table 4: Overview of molecular and polymer representations referenced in this work.

B Pre-training Dataset

We characterize the chemical diversity of our dataset of approximately 210 million polymers (total count = 210,645,750) that we use for pre-training polyBART. Polymers are categorized based on the presence of 40 functional groups. Classification is performed by analyzing the entire polymer structure, allowing a single polymer to be assigned to multiple categories.

Reaction Class	Percentage
Ether	97.2
Amide	72.5
Amine	67.1
Acetal	34.5
Halogenated	33.8
Ester	23.7
Thioether	23.6
Allyl	21.2
Hydroxyl	17.7
Sulfoxide	16
Urea	11.4
Imide	8.3
Nitrile	6
Carboxylic acid	5.2
Urethane	4.9

Table 5: Top 15 polymer classes in the pretraining data.

C Architecture and Training Setup

Configuration Parameter	polyBART _{small}	polyBART _{large}
d_{model}	256	1024
Encoder Layers	2	12
Decoder Layers	2	12
Encoder FFN Dim.	256	4096
Decoder FFN Dim.	256	4096
Encoder Attention Heads	4	16
Decoder Attention Heads	4	16

Table 6: Model architecture configurations for polyBART_{small} and polyBART_{large} as following SELFIES-TED_{small} and SELFIES-TED_{large}.

The loss function for training is given by:

$$\mathcal{L}_{\text{denoise}} = - \sum_{t=1}^T \log P(y_t \mid y_{<t}, \tilde{x}; \theta)$$

where y_t is the t -th token in the output sequence y , $y_{<t}$ is the sequence of tokens before position t , $\tilde{x}$ is the corrupted version of the original input, x , θ represents the model parameters, and $P(y_t \mid y_{<t}, \tilde{x}; \theta)$ is the model’s predicted probability of the token y_t conditioned on the previously generated tokens and the corrupted input.

polyBART_{small} is trained on ~200 million PSELFIES for 6 epochs using a per-device batch size of 128. Due to computational limitations,

polyBART_{large} is trained on 25 million samples for 5 epochs with a per-device batch size of 64. All training runs are conducted on 2 NVIDIA L40S GPUs.

D Property Datasets

We evaluate polyBART on T_g , T_d , T_m , E_{gc} , E_{gb} , E_{ea} , and E_i . Histograms of the value distributions for each property are shown below. Thermal properties are obtained from PolyInfo at <https://polymer.nims.go.jp/>. Electronic properties are accessible through the polyVERSE repository (<https://github.com/Ramprasad-Group/polyVERSE/tree/main>) and the Khazana polymer database (<https://khazana.gatech.edu/dataset/>).

D.1 Thermal Properties

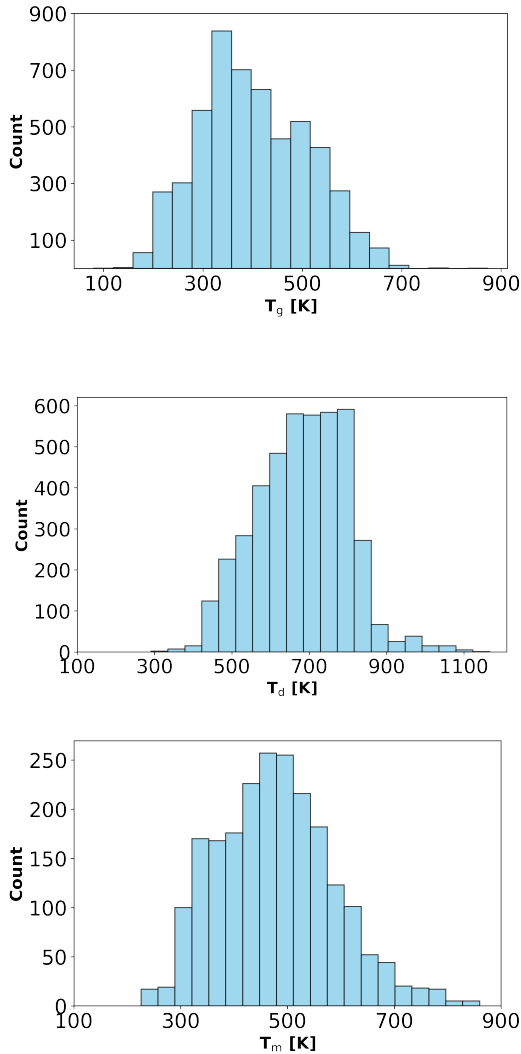


Figure 5: Distribution of thermal properties.

D.2 Electronic Properties

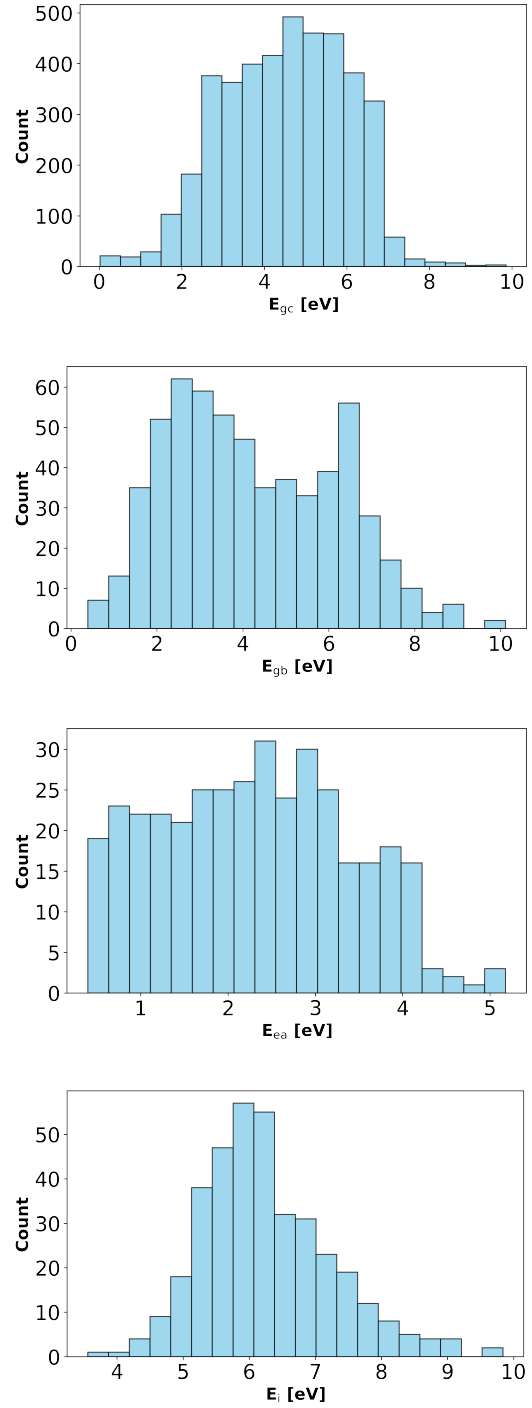


Figure 6: Distribution of electronic properties.

E Comparison with LLaMA

To enable a comparison between polyBART and an LLM, we evaluate both approaches on the task of thermal property prediction, using the 8B-parameter LLaMA-3 model as the representative LLM baseline. Because the available electronic-

property datasets are comparatively small, we limit this comparison to thermal properties.

We fine-tune LLaMA to follow a simple question-answer protocol that maps SELFIES to scalar property values. Specifically, each training instance is formatted as:

Prompt: What is the {property name} of the polymer with SELFIES: {selfies}?

Response: {property value}

We fine-tune LLaMA with LoRA ($r = 8$, $\alpha = 8$) and early stopping. At inference time, we employ deterministic decoding (temperature = 0, no sampling) to produce a single numeric answer.

F Optimal Noise Level

To identify the optimal level of noise for property-conditioned generation, we analyze the effect of varying noise levels on two metrics: validity and novelty. The following plots summarize these metrics for polyBART across different noise levels. Based on our observations we choose a noise level of 0.75 for polyBART_{small} and 1.75 for polyBART_{large}.

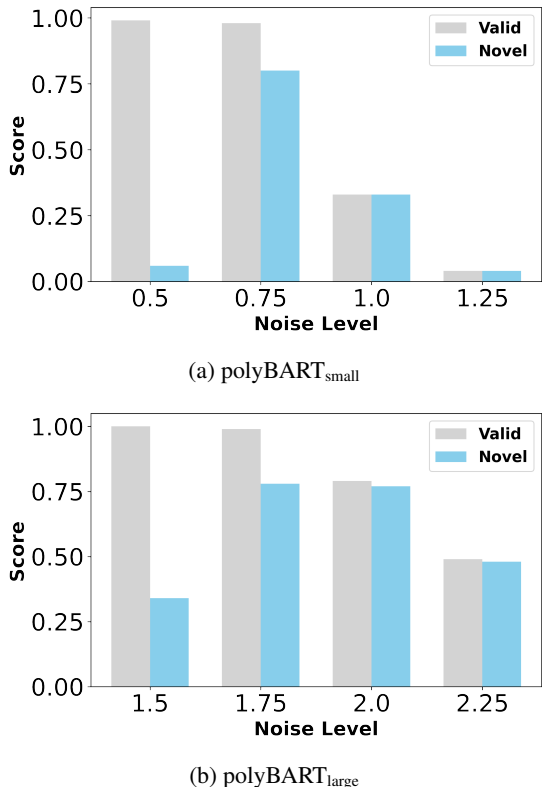


Figure 7: Validity and novelty across noise levels.

G Metrics

G.1 FCD

FCD is computed using molecular representations extracted from the penultimate layer activations of the ChemNet model. Assuming these activations follow multivariate Gaussian distributions, the mean and covariance is calculated for both generated and input molecules. If the generated distribution G has mean μ_G and covariance Σ_G , and the input distribution I has mean μ_I and covariance Σ_I , FCD between these two is given by:

$$\text{FCD}(G, I) = \|\mu_G - \mu_I\|_2^2 + \text{Tr}(\Sigma_G + \Sigma_I - 2(\Sigma_G \Sigma_I)^{1/2})$$

G.2 IntDiv_p

IntDiv_p is computed using $T(m_1, m_2)$, which denotes the Tanimoto similarity between molecules m_1 and m_2 . The summation is performed over all pairs of molecules in the set of generations G , and the result is normalized by $|G|^2$, the total number of pairwise comparisons. The parameter p controls the order of the mean. IntDiv_p is given by:

$$\text{IntDiv}_p(G) = 1 - \left(\frac{1}{|G|^2} \sum_{m_1, m_2 \in G} T(m_1, m_2)^p \right)^{\frac{1}{p}}$$

H Threshold Selection for SA Scores

We assess the distribution of SA scores within the experimental T_g dataset. As shown in Figure 12, majority of the polymers exhibit SA scores of 6 or lower, suggesting that a cutoff of 6 effectively identifies polymers that are sufficiently easy to synthesize. Hence, we adopt this threshold for our computational experiments on chemical structure generation.

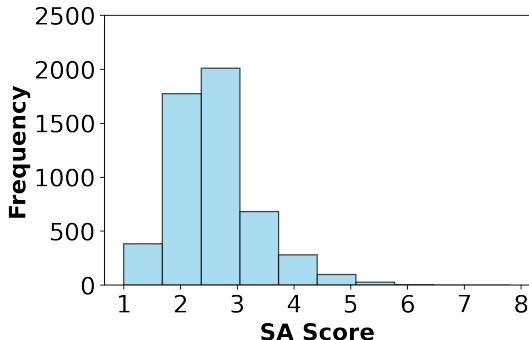


Figure 8: SA score distribution.

I Latent Space Visualization

Figure 9 shows the property-conditioned latent space visualizations for polyBART_{small} and

polyBART_{large}, with each plot highlighting regions in the embedding space associated with polymers exhibiting high T_g and E_{gc} values. The latent representations are obtained by encoding polymer structures with each model's encoder and projecting the resulting embeddings into two dimensions using t-SNE. These visualizations reveal how different models structure the latent space with respect to property-relevant features. A gradient in property values across the latent space suggests that the model has learned a meaningful representation in which polymer embeddings are aligned with their associated properties.

J Experimental Procedure

To synthesize poly(amic acid) (PAA), 3,3'-diaminodiphenylmethane (31 mg, 0.16 mmol) is added to a three-necked round-bottom flask, followed by the addition of *N,N*-dimethylacetamide (DMAc, 250 μ L) as the solvent. The mixture is stirred until the diamine was completely dissolved. Subsequently, an equimolar amount of 3,3',4,4'-benzophenonetetracarboxylic dianhydride (50 mg, 0.16 mmol) is slowly added to the solution under a nitrogen atmosphere. The reaction temperature is maintained between 40°C and 50°C, and the mixture is stirred overnight to obtain a homogeneous PAA solution. The resulting solution is slowly poured into deionized water under vigorous stirring, leading to the formation of a white to pale-yellow precipitate. The precipitate is collected by filtration and washed thoroughly with acetone multiple times to remove residual solvent. The purified PAA is air-dried at room temperature and then vacuum-dried at 60°C until a constant weight is achieved.

Thermal imidization of the obtained PAA is conducted to afford the corresponding polyimide (PI). The dried PAA is placed in a flask under vacuum, and the temperature is gradually increased in a step-wise manner to 150°C, 200°C, and finally 300°C, holding each temperature for approximately 1 hour. After cooling to room temperature under vacuum, the resulting PI (~54 mg, 71%) is collected for further characterization.

Thermogravimetric analysis (TGA) was performed using a Pyris 1 TGA (PerkinElmer) at a heating rate of 10°C/min. Differential Scanning Calorimetry (DSC) was carried out on a DSC 3+ STARe system (Mettler Toledo), with both heating and cooling rates set to 10°C/min.

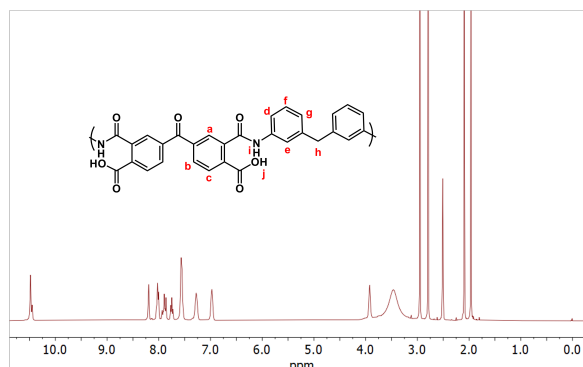


Figure 10: ^1H NMR spectrum of PAA in d_6 -DMSO.

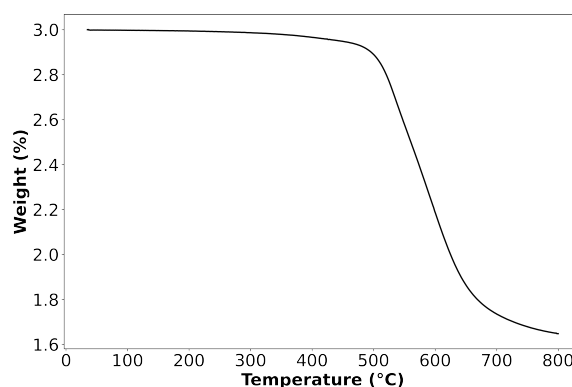


Figure 11: TGA curve of PI.

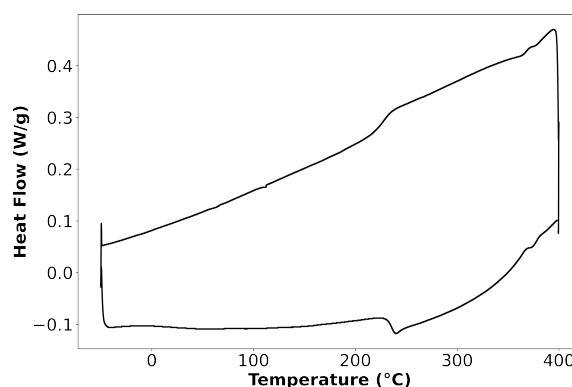


Figure 12: DSC curve of PI.

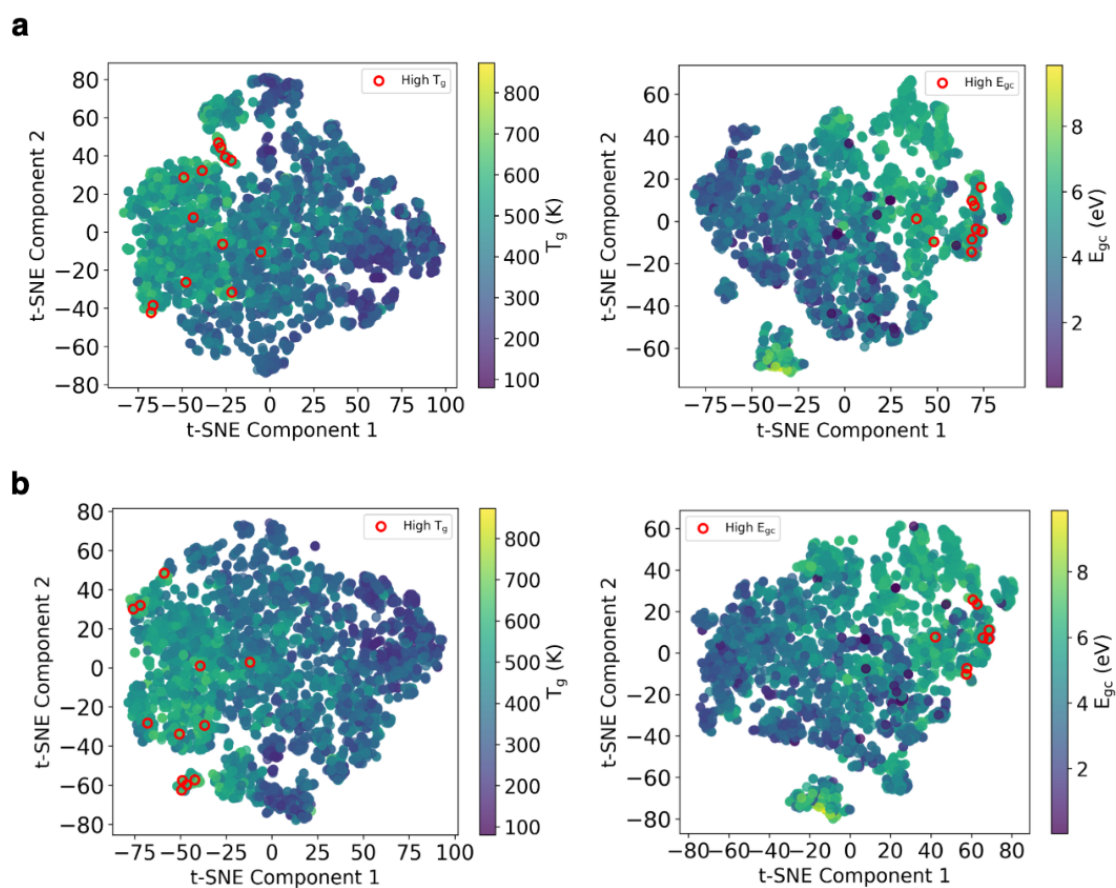


Figure 9: Latent space projections for (a) polyBART_{small} and (b) polyBART_{large} highlighting polymers exhibiting high T_g and E_{gc} values.