TransBERT Polymer Informatics: A Fusion of Transformer Language Modeling and Machine-Driven Chemistry for Accelerated Property Predictions

BHAUMIK TYAGI[1], PRATHAM TANEJA[2], AKSHITA GUPTA[3] , DAAMINI BATRA[4] and KESHAV CHANDRA[5]

[1]*Jr. Research Scientist, Delhi, India*
[2]*Graduate Student, (Electronics and Communication Engineering), ADGITM, Delhi, India*
[3,4,5] *Undergraduate Student, (Information Technology), ADGITM, Delhi, India*

Abstract—This research introduces a pioneering framework named TransBERT that capitalizes on the capabilities of two sophisticated language models, TransPolymer and polyBERT, to comprehensively advance the polymer informatics field. TransPolymer, a Transformer-based language model, predicts polymer properties by leveraging self-attention mechanisms. The model employs a polymer tokenizer imbued with chemical awareness, facilitating the extraction of meaningful representations from polymer sequences. Moreover, TransPolymer benefits from rigorous pretraining on extensive unlabeled datasets through Masked Language Modeling, underscoring the pivotal role of self-attention in effectively modeling polymer sequences. In conjunction with TransPolymer, polyBERT contributes a fully automated polymer informatics pipeline designed to expedite the identification of application-specific polymer candidates with heightened speed and accuracy. Drawing inspiration from Natural Language Processing concepts, polyBERT operates as a chemical linguist, treating the chemical structure of polymers as a unique language. The pipeline integrates a polymer chemical fingerprinting capability and a multitask learning approach to map polyBERT fingerprints to diverse polymer properties effectively. Notably, polyBERT outperforms existing polymer property prediction methods based on manually crafted fingerprint schemes by achieving a remarkable two orders of magnitude increase in speed while maintaining high accuracy and integrating TransPolymer and polyBERT results in a robust computational tool poised to propel the fields of polymer design and structure-property relationship understanding. This combined framework strategically harnesses the strengths of Transformer models and machine-driven informatics, offering unparalleled efficiency in the prediction and identification of polymer properties. This synergistic approach holds significant promise for scalable deployment, including applications in cloud infrastructures, thereby making substantial contributions to the advancement of polymer science and informatics.

Keywords— Polymer Informatics, TransPolymer, PolyBERT, NLP, Machine-driven informatics

I. INTRODUCTION

The establishment of rational representations that effectively map polymers into a continuous vector space is imperative for the successful application of machine learning tools in polymer property prediction. The precision and efficiency of property prediction play a pivotal role in the strategic design of polymers for diverse applications, spanning from polymer electrolytes [1] to organic optoelectronics [2], energy storage [3], and various other fields [4]. To enhance the predictive capabilities in polymer-related tasks, we introduce fingerprints (FPs), a proven and effective approach derived from molecular machine-learning models [5]. Recent advancements in deep neural networks (DNNs) have revolutionized polymer property prediction by enabling the direct learning of expressive representations from data, leading to the generation of deep fingerprints. This innovative approach eliminates the reliance on manually engineered descriptors [6]. While Graph Neural Network (GNN)--based models have demonstrated significant progress in polymer property prediction, they necessitate explicit knowledge of structural and conformational information, which can be computationally or experimentally expensive to acquire. In the realm of polymer informatics pipelines, a crucial step involves the conversion of polymer chemical structures into numerical representations commonly referred to as fingerprints, features, or descriptors. This research underscores the evolving landscape of polymer property prediction methodologies, emphasizing the shift towards data-driven approaches facilitated by deep learning techniques. The integration of fingerprints and the elimination of manual descriptor engineering mark a significant leap forward, streamlining the prediction process and contributing to the advancement of polymer science and informatics.

Historically, prior methodologies for fingerprinting in polymer research [17] have relied on

cheminformatics tools to numerically encode essential chemical and structural features. While these handcrafted fingerprinting approaches are rooted in valuable intuition and accumulated experience, their development is characterized by a laborious and intricate process. This involves complex computations, consuming a substantial portion of time during both model training and inference phases. Furthermore, the resultant fingerprints often lack generalizability across all polymer chemical classes, necessitating ad hoc additions to the feature catalogue when encountering new classes. The reliance on handcrafted fingerprints introduces inherent challenges within machine learning (ML) pipelines, particularly in the exploration of novel polymer chemical classes. Such pipelines, utilizing manually engineered fingerprints, are susceptible to errors and may encounter difficulties accommodating diverse chemical structures. Additionally, these handcrafted approaches present obstacles to the realization and deployment of fully machine-driven pipelines, which are essential for achieving scalability in cloud computing and high-throughput environments. This research underscores the limitations of traditional handcrafted fingerprinting methodologies, emphasizing the need for more scalable and adaptable approaches in the context of modern polymer informatics. The transition towards fully machine-driven pipelines is essential for overcoming the challenges associated with diverse polymer chemical classes and facilitating seamless integration into scalable computing environments.

To address the aforementioned constraints, a promising strategy involves the substitution of manual fingerprinting methodologies with machine-crafted counterparts, particularly those generated through the application of "Transformer" technology. Transformers, a recent innovation originating from Natural Language Processing (NLP), have swiftly emerged as the benchmark in machine learning language modeling [18]. This study introduces a novel paradigm where Simplified Molecular-Input Line-Entry System (SMILES) [19] strings, commonly employed for polymer representation, serve as the foundational "chemical language" for polymers. The approach involves the utilization of millions of Polymer SMILES (PSMILES) strings to train a language model named polyBERT. This model is designed to transcend beyond conventional fingerprinting methods, evolving into an expert—a linguist—proficient in deciphering the intricate chemical language specific to polymers.

II. LITERATURE REVIEW

Recurrent Neural Network (RNN)-based models, commonly employed for encoding chemical knowledge from polymer sequences, often face limitations in competitiveness. This is attributed to their reliance on previous hidden states for capturing dependencies between words, resulting in information loss as the model progresses to deeper steps. In contrast, the transformative impact of Transformer models on natural language processing (NLP) tasks, as evidenced by their exceptional performance in recent years [7], has prompted a reevaluation of their application in chemistry and materials science.

The Transformer and its variants, noted for their attention mechanism, have demonstrated a paradigm shift in NLP tasks. This architecture excels in capturing relationships between tokens in a sequence without relying on past hidden states. Notable Transformer-based models such as BERT [8], RoBERTa [9], GPT [10], ELMo [11], and XLM [12] have emerged as effective pretraining methods through self-supervised learning, enhancing representations derived from unlabeled texts and subsequently improving performance across diverse downstream tasks. This paper explores the potential of Transformer models in the context of chemistry and materials science, highlighting their unique ability to overcome the limitations associated with traditional RNN-based approaches. The transformative impact of attention mechanisms, coupled with the success of various Transformer-based pretraining methods, underscores the promising avenue these models present for advancing the understanding and application of language-based approaches in chemical and materials informatics.

The application of Transformer models in predicting the properties of small organic molecules has been demonstrated [13]. However, when extended to sequence models for polymers, a notable challenge emerges due to the inherent scarcity of readily available and well-labeled data. This scarcity is exacerbated by the labor-intensive nature of the characterization process in laboratory settings, compounded further by limited accessibility to certain polymer data sources [14]. The utilization of Transformer models, exemplified by TransPolymer, proves advantageous in encoding chemical information about the internal interactions of polymers and influential factors governing polymer properties. The observation of attention scores through visualization provides empirical evidence of TransPolymer's capacity to learn generalizable

features, thereby facilitating their transferability to the prediction of polymer properties. This ability holds significant implications for polymer design, emphasizing the broader applicability of TransPolymer in addressing challenges related to limited and inaccessible polymer data sources.

Recent investigations [20] have underscored the advantageous utilization of Transformers within the molecular chemical space. Notably, Wang et al. [21] demonstrated the efficacy of training a BERT [22] model, a widely adopted general language model, using a dataset of molecule Simplified Molecular-Input Line-Entry System (SMILES) strings. Leveraging BERT's latent space representations as molecular fingerprints, the authors observed superior performance compared to other fingerprinting methods, including those based on unsupervised recurrent neural networks and graph neural networks. A parallel effort by Schwaller et al. [23] introduced a Transformer model for predicting retrosynthesis pathways of molecules, surpassing established algorithms in the field of reaction prediction. In a recent study, Xu et al. [24] harnessed a RoBERTa model, an evolution of the BERT Transformer, for polymer property predictions. Their approach involved a two-step process, commencing with the pretraining of the RoBERTa model through unsupervised learning on a dataset of 5 million polymers. Subsequently, a fine-tuning step, conducted through supervised training, enabled direct predictions of polymer properties. Additionally, alternative neural network architectures, specifically graph neural networks [25], have been applied to both the molecule and polymer chemical spaces in prior research. In contrast to Transformers, graph neural networks represent atoms as nodes and bonds as edges within a graph, capturing immediate and extended connectivities between atoms. Unlike Transformers, graph neural networks do not rely on Polymer SMILES (PSMILES) strings but necessitate an initial set of feature vectors (such as atom types, implicit valence, etc.) assigned to each node.

III. METHODOLOGY

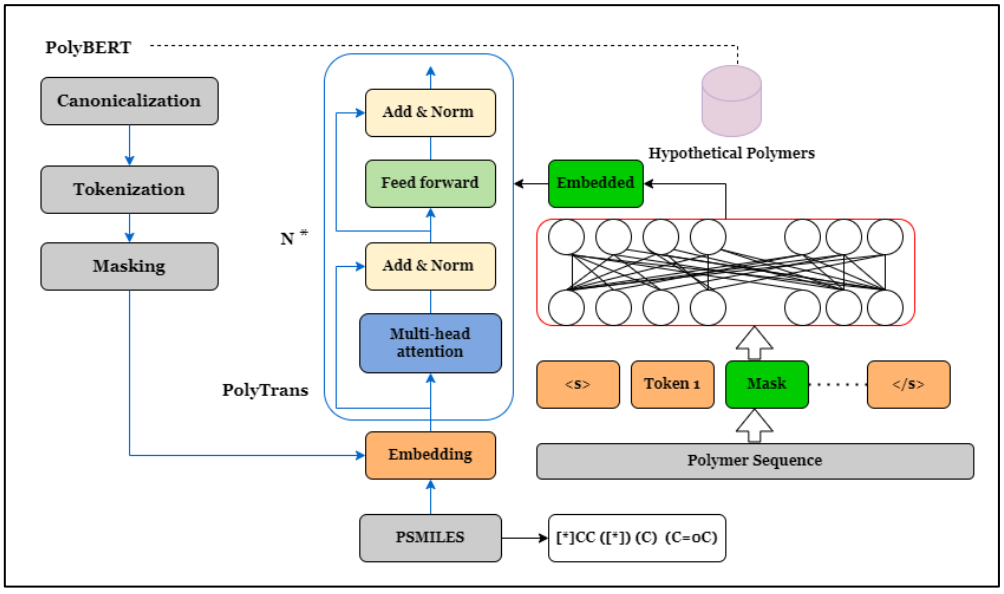


Fig. 1 | Polymer informatics with TransBERT. a Prediction pipeline.

The left pipeline shows the prediction using handcrafted fingerprints using cheminformatics tools, while the right pipeline (present work) portrays a fully end-to-end machine-driven predictor using polyBERT. Illustration of the pretraining (left) and finetuning (right) phases of TransPolymer. The model is pretrained with Masked Language Modeling to recover original tokens, while the feature vector corresponding to the special token '⟨s⟩' of the last hidden layer is used for prediction when finetuning. Within the TransPolymer block, lines of deeper color and larger width stand for higher attention scores.

Transformer based encoder:

Unlike RNN-based models which encoded temporal information by recurrence, Transformer uses self-attention layers instead. The attention mechanism used in Transformer is named Scaled Dot-Product Attention, which maps input data into three vectors: queries (Q), keys (K), and values (V). The attention is computed by first computing

the dot product of the query with all keys, dividing each by $\sqrt{d_k}$ for scaling where d_k is the dimension of keys, applying the SoftMax function to obtain the weights of values, and finally deriving the attention. The dot product between queries and keys computes how closely aligned the keys are with the queries. Therefore, the attention score can reflect how closely related the two embeddings of tokens are. The formula of Scaled Dot-Product Attention can be written as:

$$\text{Attention}\ (Q, K, V) = \text{softmax}\ (QK^T/\sqrt{d_k})\ V$$

Multi-head attention is performed instead of single attention by linearly projecting Q, K, and V with different projections and applying the attention function in parallel. The outputs are concatenated and projected again to obtain the results. In this way, information from different subspaces could be learned by the model.

TransPolymer framework: Our TransPolymer framework consists of tokenization, Transformer encoder, pretraining, and finetuning. Each polymer data is first converted to a string of tokens through tokenization. Polymer sequences are more challenging to design than molecule or protein sequences as polymers contain complex hierarchical structures and compositions. For instance, two polymers that have the same repeating units can vary in terms of the degree of polymerization.

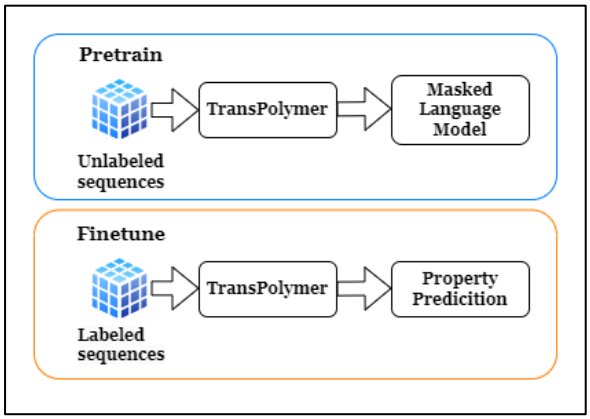


Fig. 2 | The whole TransPolymer framework with a pre-train-finetune pipeline.

The benchmark, whose size is around 1M, was built by Ma et al. by training a generative model on polymer data collected from the PolyInfo database [15]. The generated sequences consist of monomer SMILES and '*' signs representing the polymerization points. The ~1M database was demonstrated to cover similar chemical space as PolyInfo but populate space where data in PolyInfo are sparse. Therefore, the database can serve as an important benchmark for multiple tasks in polymer informatics. To finetune the pretrained TransPolymer, ten datasets are used in our experiments which cover various properties of different polymer materials, and the distributions of polymer sequence lengths vary from each other.

The performance of our pretrained TransPolymer model on ten property prediction tasks is illustrated below. We use root mean square error (RMSE) and R2 as metrics for evaluation. For each benchmark, the baseline models and data splitting are adopted from the original literature. we develop long shortterm memory (LSTM), another widely used language model, as well as unpretrained TransPolymer trained purely via supervised learning as baseline models in all the benchmarks. $\text{TransPolymer}_{\text{unpretrained}}$ and $\text{TransPolymer}_{\text{pretrained}}$ denote unpretrained and pretrained TransPolymer, respectively.

IV. RESULTS

The results of TransPolymer and baselines on PE-I are illustrated in Table 2.

Table 1. Performance of TransPolymer and baseline models on PE-I

Model	**Train RMSE**	**Test RMSE**	**Train R^2**	**Test R^2**
$\text{TransPolymer}_{\text{unpretatined}}$	0.90	1.03	0.71	0.32
LSTM	1.05	1.46	0.69	-0.27
$\text{TransPolymer}_{\text{pretatined}}$	0.22	0.69	0.99	0.70

$\text{TransPolymer}_{\text{pretrained}}$, which achieves the lowest RMSE of 0.69 and highest R2 of 0.99 on the average of cross-validation sets, exhibits better generalization.

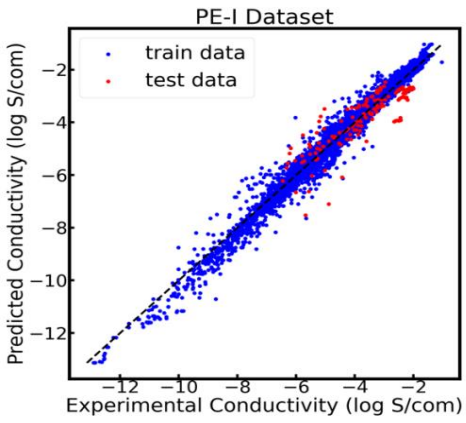


Fig. 3 | Scatter plots of predicted values by $TransPolymer_{pretrained}$ and baseline model on PE1 dataset

Table 2. Performance of TransPolymer and baseline models on OPV

Model	Train RMSE	Test RMSE	Train R^2	Test R^2
$TransPolymer_{unpretained}$	1.92	2.12	0.37	0.20
LSTM	2.37	2.36	-0.02	0.01
$TransPolymer_{pretatined}$	1.20	1.93	0.75	0.33

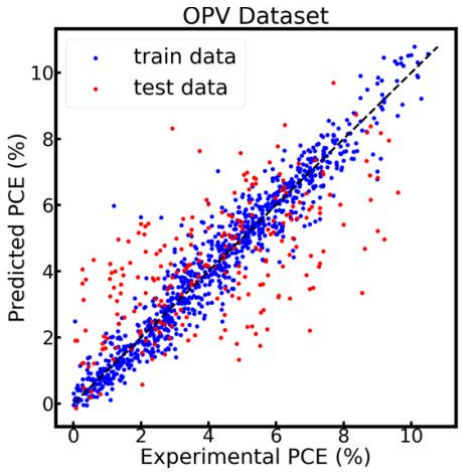
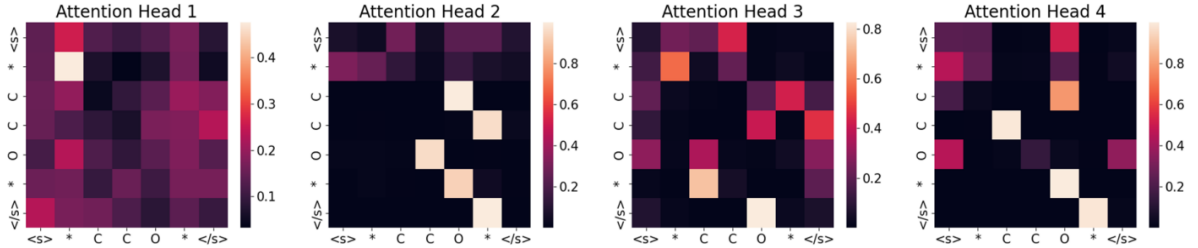
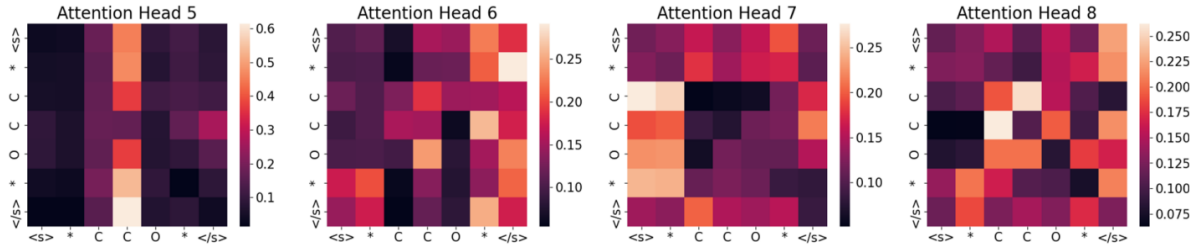


Fig. 4 | Scatter plots of predicted values by $TransPolymer_{pretrained}$ and baseline model on OPV dataset

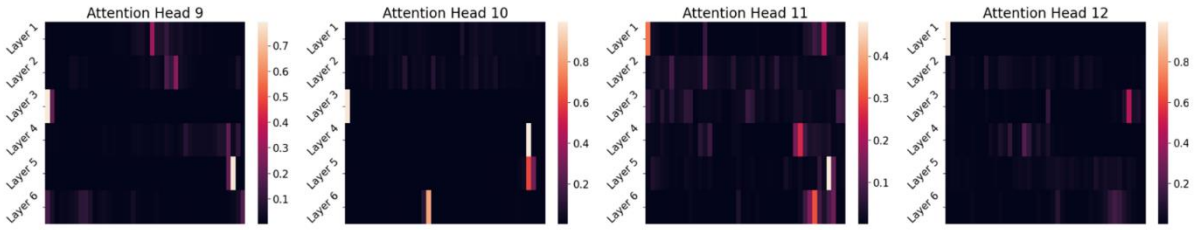
This research introduces TransPolymer, a Transformer-based model with Masked Language Modeling (MLM) pretraining, positioned as an advanced solution for accurate and efficient polymer property prediction. The proposed model employs a meticulously designed polymer tokenization strategy to effectively map polymer instances to sequences of tokens. Data augmentation strategies are implemented to augment the available data, enhancing the model's capabilities in representation learning. TransPolymer undergoes a two-step training process, commencing with MLM pretraining on approximately 5 million unlabeled polymer sequences, followed by fine-tuning on diverse downstream datasets. This comprehensive training approach results in TransPolymer outperforming all baselines and unpretrained versions. The superior model performance is attributed to the impact of pretraining with a substantial amount of unlabeled data, fine-tuning Transformer encoders, and data augmentation for expanding the data space. Attention scores from hidden layers in TransPolymer offer empirical evidence of the model's efficacy in learning representations with chemical awareness and identifying influential tokens in final prediction results. The study anticipates that TransPolymer, with its desirable model performance and exceptional generalization ability even with limited labeled downstream data, holds potential as a solution for predicting newly designed polymer properties and guiding polymer design. The pretrained TransPolymer is envisioned to be applied in an active-learning-guided polymer discovery framework, contributing to virtual screening of the polymer space, recommending potential candidates based on model predictions, and updating through learning on data from experimental evaluation. Furthermore, TransPolymer exhibits outstanding performance on copolymer datasets compared to existing baseline models, thereby shedding light on the exploration of copolymers. Although the primary focus of this paper centers on regression, TransPolymer is positioned to pave the way for promising (co)polymer discovery frameworks.



(A) Attention scores in the first hidden layer.



(B) Attention scores in the last hidden layer.



(C) Visualization of attention scores from finetuned TransBERT.

Fig. 5 | Visualization of attention scores from pretrained TransPolymer.

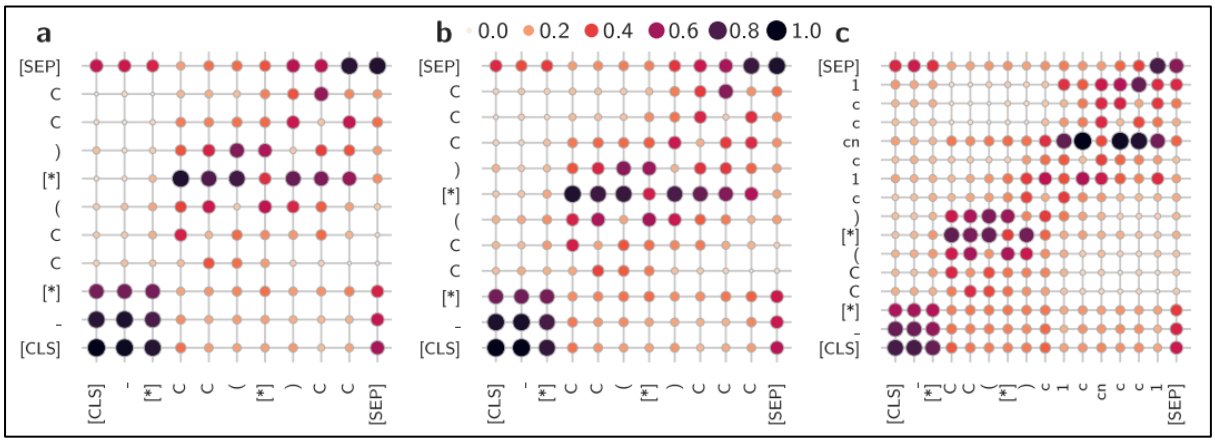


Fig. 6 | Attention maps and neuron activation for three polymers. Panels a–c shows the normalized attention maps summed over all 12 attention heads and 12 encoders of TransBERT.

This study presents a highly adaptable, expeditious, and precise polymer informatics pipeline designed for seamless scalability on cloud hardware, specifically tailored for high-throughput screening of extensive polymer spaces. At the core of this pipeline is polyBERT, a Transformer-based Natural Language Processing (NLP) model engineered for the nuances of polymer chemical language. Trained on a dataset comprising 100 million hypothetical polymers, polyBERT forms the cornerstone of an informatics pipeline that delivers polymer representations and predicts polymer properties at speeds two orders of magnitude faster than the most effective pipeline relying on manually crafted fingerprints. The enormity of the polymer universe, constrained by current limitations in experimentation, manufacturing techniques, resources, and economic considerations, necessitates novel approaches for exploration. Considering various polymer types, including homo-polymers, copolymers, and polymer blends, alongside unexplored chemistries, additives, and processing conditions, the potential diversity within the polymer universe is limitless. However, the exploration of this vast space, enabled by property predictions, is currently hindered by prediction speed. The accurate prediction of 29 properties for 100 million hypothetical polymers within a reasonable timeframe underscores polyBERT's role as an enabler for extensive exploration of the vast polymer universe at scale.

TransBERT not only accelerates polymer informatics pipelines, surpassing state-of-the-art approaches by a factor of 100, but also maintains accuracy comparable to slower handcrafted fingerprinting methods. Leveraging Transformer-based Machine Learning (ML) models originally developed for Natural Language Processing, TransBERT fingerprints emerge as dense and chemically pertinent numerical representations facilitating precise measurement of polymer similarity. These fingerprints find application in various polymer informatics tasks, including property predictions, polymer structure predictions, and ML-based synthesis assistance. The potential of TransBERT fingerprints to replace handcrafted fingerprints in accelerating polymer informatics pipelines is significant. Additionally, TransBERT holds promise in directly designing polymers based on fingerprints, a prospect that entails retraining and

structural updates to TransBERT, marking a direction for future work.

V. CONCLUSION

This research presents two innovative approaches, TransPolymer and polyBERT, contributing to the advancement of polymer informatics and property prediction. TransPolymer, a Transformer-based model, employs Masked Language Modeling (MLM) pretraining on approximately 5 million unlabeled polymer sequences, demonstrating superior performance in accurate and efficient polymer property prediction. Through a well-designed polymer tokenization strategy and data augmentation, TransPolymer excels in representation learning, outperforming baselines and $un_{pretrained}$ versions. Attention scores from hidden layers provide insights into the model's capacity to learn chemical representations and influential factors in polymer properties. Concurrently, polyBERT, a Transformer-based Natural Language Processing (NLP) model, serves as a critical component in a generalizable, ultrafast, and accurate polymer informatics pipeline. Trained on 100 million hypothetical polymers, polyBERT facilitates predictions of polymer properties at speeds two orders of magnitude faster than pipelines relying on handcrafted fingerprints, maintaining high accuracy. The scalability of the polyBERT-based informatics pipeline on cloud hardware enables high-throughput screening of extensive polymer spaces. The collective superior performance of TransPolymer and polyBERT highlights their potential in guiding polymer design and exploration. TransPolymer's application in an active-learning-guided polymer discovery framework and its notable performance on copolymer datasets suggest promising avenues for future research. polyBERT, with its dense and chemically pertinent numerical representations of polymers, accelerates polymer informatics pipelines by replacing handcrafted fingerprints. The study concludes by emphasizing the potential of polyBERT fingerprints in various polymer informatics tasks and its role in future advancements, such as direct polymer design based on fingerprints through retraining and structural updates.

REFERENCES

[1] Wang, Y. et al. Toward designing highly conductive polymer electrolytes by machine learning-assisted coarse-grained molecular dynamics. Chem. Mater. 32, 4144–4151 (2020).

[2] St. John, P. C. et al. Message-passing neural networks for high-throughput polymer screening. J. Chem. Phys. 150, 234111 (2019).

[3] Luo, H. et al. Core–shell nanostructure design in polymer nanocomposite capacitors for energy storage applications. ACS Sustain. Chem. Eng. 7, 3145–3153 (2018).

[4] Liang, J., Xu, S., Hu, L., Zhao, Y. & Zhu, X. Machine-learning-assisted low dielectric constant polymer discovery. Mater. Chem. Front. 5, 3823–3829 (2021).

[5] Mannodi-Kanakkithodi, A. et al. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. Mater. Today 21, 785–796 (2018).

[6] Chen, L. et al. Polymer informatics: current status and critical next steps. Mater. Sci. Eng. R. Rep. 144, 100595 (2021).

[7] Vaswani, A. et al. Attention is all you need. Adv. Neural. Inf. Process. Syst. 30, (2017).

[8] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 4171–4186 (2019).

[9] Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).

[10] Brown, T. et al. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901 (2020).

[11] Peters, M. E., Neumann, M., Zettlemoyer, L. & Yih, W.-t. Dissecting contextual word embeddings: architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 1499–1509 (2018).

[12] Conneau, A. & Lample, G. Cross-lingual language model pretraining. Adv. Neural. Inf. Process. Syst. 32, (2019).

[13] Honda, S., Shi, S. & Ueda, H. R. Smiles transformer: a pre-trained molecular fingerprint for low data drug discovery. Preprint at https://arxiv.org/abs/1911.04738 (2019).

[14] Persson, N., McBride, M., Grover, M. & Reichmanis, E. Silicon valley meets the ivory tower: searchable data repositories for experimental nanomaterials research. Curr.

Opin. Solid State Mater. Sci. 20, 338–343 (2016).

[15] Ma, R. & Luo, T. Pi1m: a benchmark database for polymer informatics. J. Chem. Inf. Model 60, 4684–4690 (2020).

[16] Reis, M. et al. Machine-learning-guided discovery of 19f mri agents enabled by automated copolymer synthesis. J. Am. Chem. Soc. 143, 17677–17689 (2021).

[17] Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative structure-property relationship modeling of diverse materials properties. Chem. Rev. 112, 2889–2919 (2012).

[18] Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. J. Cheminf. 10, 4 (2018).

[19] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model.28, 31–36 (1988).

[20] Chithrananda, S., Grand, G., Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv (2020), https://doi.org/10.48550/arXiv.2010.09885.

[21] Wang, S., Guo, Y., Wang, Y., Sun, H., Huang, J. SMILES-BERT. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA; pp 429–436, (2019). https://doi.org/10.1145/3307339. 3342186.

[22] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv (2018), https://doi.org/10.48550/arXiv.1810.04805.

[23] Schwaller, P. et al. Molecular transformer: A model for uncertaintycalibrated chemical reaction prediction. ACS Central Sci. 5, 1572–1583 (2019).

[24] Xu, C., Wang, Y., Farimani, A. B. TransPolymer: a transformer-based language model for polymer property predictions. arXiv (2022), https://doi.org/10.48550/arXiv.2209.01307.

[25] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., Dahl, G. E. Neural message passing for quantum chemistry. arXiv (2017), https://doi.org/10.48550/arXiv.1704.01212.