A Large Encoder-Decoder Polymer-Based Foundation Model

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Nathaniel Park
IBM Research Almaden
San Jose, CA, USA
npark@us.ibm.com

Emilio Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Victor Shirasuna
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
victor.shirasuna@ibm.com

Abstract

Representation systems for polymers are a constant issue in the development of deep-learning models for polymer property prediction, necessitating a balance between structural accuracy, interpretability, and interoperability to achieve utility across prediction tasks. To facilitate this, we introduce a serialized polymer graph (SPG) notation and SPG-TED$_{289M}$, a SPG-based foundation model for polymers, which has been pre-trained on a carefully curated dataset of 1 million SPG samples. To better handle the unique characteristics of SPG, we extended the tokenization process, resulting in a vocabulary of 2,407 distinct tokens. We evaluated the SPG-TED$_{289M}$ model's performance across a range of tasks including copolymer electron affinity and ionization potential, polymer membrane properties, multi-task learning, refractive index prediction, ionic conductivity, gas permeability, and glass transition temperature. The model demonstrated state-of-the-art performance in most of these areas, achieving results on par with specialized models designed for specific tasks. This indicates that SPG-TED$_{289M}$, with minimal fine-tuning, can adapt effectively to complex polymer-related tasks, showcasing its robustness and versatility as a foundation model. The SPG-TED$_{289M}$ model provides significant flexibility and scalability, making it a valuable tool for various applications in polymer science.

1 Introduction

The creation of generalizable predictive models for polymer properties is a challenging yet critical endeavor to accelerate the development of high-performance polymeric materials. The ability to accurately predict polymer properties based solely on their respective structural and architectural features is imperative to guide experimental design choices. This is particularly relevant for applications leveraging automated or autonomous experimentation platforms,[1] where active learning or reinforcement learning algorithms may select new candidates based on both on predicted properties and feasibility within the automated system. However, both the stochastic nature of polymers and myriad possible architectures makes it difficult to represent their structure accurately and discretely—a necessity for construction of meaningful models for property prediction. Such difficulties in polymer representation are frequently circumvented by modeling polymer structures as discrete SMILES strings with connectivity between repeat units denoted with an asterisk, a notation commonly referred to as PSMILES.[2, 3] These text-based representations are either used directly as input to predictive models[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] or converted to another format, such as

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

images[9, 13] or graphs,[13, 15, 16, 17] prior to ingestion. Models built using PSMILES and other representations have been demonstrated to be effective in predicting a single or handful of polymer properties,[6, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21] however, very few have been shown to generalize across a range physical and chemical properties.[3, 4, 5, 7] Moreover, the majority of predictive models for polymers are overwhelmingly focused on homopolymers, frequently without accounting for end groups.[6, 8] For models which can accomodate for copolymer architectures, these are largely relegated to simple random copolymers, block copolymers, or alternating copolymers where the topology is already pre-encoded into the PSMILES string.[3, 11, 15, 19, 22] Finally, other SMILES-based representations such as BigSMILES have been created to address some of the limitations of PSMILES and is capable of at representing a broad range of complex architectures.[23] However, the more complex syntax makes it less interoperable with existing literature datasets overwhelmingly represented as PSMILES as well as requiring a significantly more specialized token vocabulary for chemistry language models.

To overcome limitations of existing polymer representation schemes and construct a highly effective foundation model for polymer property predictions, a straightforward representation system is needed that could accommodate both a range of polymer architectures and is easily be interoperable with existing literature data. Recently, a domain-specific programming language, termed Chemical Markdown Language (CMDL) was developed and enabled researchers to easily construct polymer graph representations—including both the connectivity between critical structural components of the polymer and their relation to experimentally measured values.[24] The CMDL polymer graph representations could then be compiled and exported, serializing the polymer graph representation to a string for downstream use in regression transformer models for property prediction and structure generation.[24, 25, 26, 27] The CMDL polymer graph representation also exploited the use of edge weights based on polymer architecture symmetry, avoiding a redundant 1:1 representation of polymer structural components in complex architectures and facilitating a more compact serialized output. Despite these advantages and demonstrated effectiveness as a chemical language model inputs, the serialized CMDL graph output was still somewhat verbose and required special modifications to the tokenizer in the regression transformer model. Hence, we surmised that revising the this representation to more closely resemble PSMILES yet keep information regarding edge connectivity between structural components would provide a new serialized polymer graph (SPG) representation that could accommodate both a broad array of polymer architectures and be easily interoperable with existing literature datasets—simplifying assembly of pre-training and benchmarking datasets.

2 Related Work

As noted above, relatively few efforts exist in the development of foundation models for polymers which can generalize across a range of prediction tasks for physical and chemical properties. One such effort is polyBERT, a transformer model which is pre-trained on 100 million hypothetical PSMILES strings—which are generated from a fragmentation and recombination of >13K previously synthesized polymers—and then fine-tuned for predictive tasks across several polymer property classes.[3] Another model, TransPolymer, was pre-trained on 5 million PSMILES augmented from the 1 million present in PI1M, a PSMILES dataset generated from the PolyInfo database.[28] The model was then fine-tuned augmented versions of common DFT benchmark datasets.[2] MMPolymer is an analogous multimodal transformer model, using both PSMILES input as well as 3D conformations to facilitate predictive tasks across several datasets.[4] Finally, SML-MT is a foundation model for polymer property prediction pre-trained on 1 billion SMILES from Enamine REAL dataset and then fine-tuned against several DFT benchmark datasets represented using PSMILES.[7]

3 Overview of the Proposed Approach

This section provides an overview of the proposed SPG-based foundation model for polymers. Here, we detail the processes of representation design, collecting, curating, and pre-processing the pre-training data, the token encoding, and SPG encoder-decoder processes. Figure 1 illustrates the general architecture of the base model.

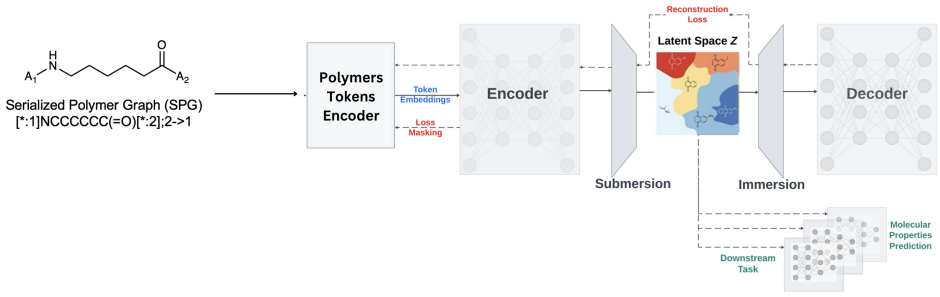

Figure 1: This figure illustrates the general architecture of the base SPG-TED$_{289M}$ model.

3.1 Polymer Representation Design

The serialization protocol for CMDL polymer graphs was adjusted to minimize introduction of new tokens to the model vocabulary and facilitate straightforward interconversion of existing benchmarking datasets. Prior non-atomic place holder characters used in CMDL polymer graphs (*R* or *Q*, (**1c**) were replaced with asterisks enclosed in brackets and numbered as shown with an example Nylon 6 homopolymer (**1d**, Figure 2A). Edge connections between structural components are denoted using the numeric labels of the attachement points following the SMILES strings (**1d**, Figure 2A). For polymers or copolymers with more structural components, each component is separated with a semicolon as opposed to a more commonly used dot separator,[15] as the semicolon may facilitate better distinction of charged components (Figure 2B). Copolymers with different architectures, such as block versus statistical copolymers (Figure 2B), but identical structural components may be distinguished based on their edge connectivity. With the random copolymer **1e**, the extra edge between the lactide and trimethylene carbonate repeat units (5 -> 4, **1e**) is present to indicate statistical bonding between the two repeat units whereas AB block copolymer **1d** lacks such a feature (Figure 2B). Finally, representation of formulations, which include both polymeric and small-molecule components, was accomplished by appending additional components with a semi-colon as shown with **1f** where the lithium bis(trifluoromethanesulfonyl)imide salt was separated from the polymer component with a semi-colon (Figure 2C).

Additional modifications were made to ensure smooth conversion of existing literature datasets using PSMILES into the SPG notation. Despite the arrow notation to indicate an edge connection between two components, no directionality is implied or assumed. Additionally, only the minimal number of edges needed to describe the connectivity are indicated in contrast to other graph-focused approaches.[15] Finally, no attempt was made to ensure the edge connection between two structural components was indeed a chemically realistic in terms of known polymer forming reactions. This is a reasonable modification as many literature datasets are homopolymers and those that contain copolymers tend to focus on a narrow set of polymerization reactions (e.g. radical polymerization, polycondensation, ring-opening polymerization, etc.).[11, 15] These modifications enabled asterisks within a SMILES string to be numbered programmatically along with a set of edge labels based on existing architecture labels for the polymer.

3.2 Pre-training Data

Collection of polymer data for pre-training suffers from two distinct problems: 1) scarcity of large, open curated sets of experimentally realized materials and 2) potential lack of structural diversity given the stringent requirements for a successful polymerization reaction placing a limit on the types of viable structural components of a polymer. In overcoming the first problem, the use of synthetic or combinatorially generated datasets is almost unavoidable to reach a sufficient volume of pre-training data. While the latter issue of improving structural diversity in the face of experimental reality is more challenging; arbitrary insertion of two or more asterisks into a SMILES string—as is frequently encountered in generative models[28, 29, 30, 31]—does not turn it into a feasible polymeric repeat unit. Additionally, synthetic datasets are known to contain non-viable polymer structures, including anti-aromatic rings or structures with unstable oxidation states for heavy atoms.[28] Hence, the pre-training data were carefully pooled from a selection of open literature

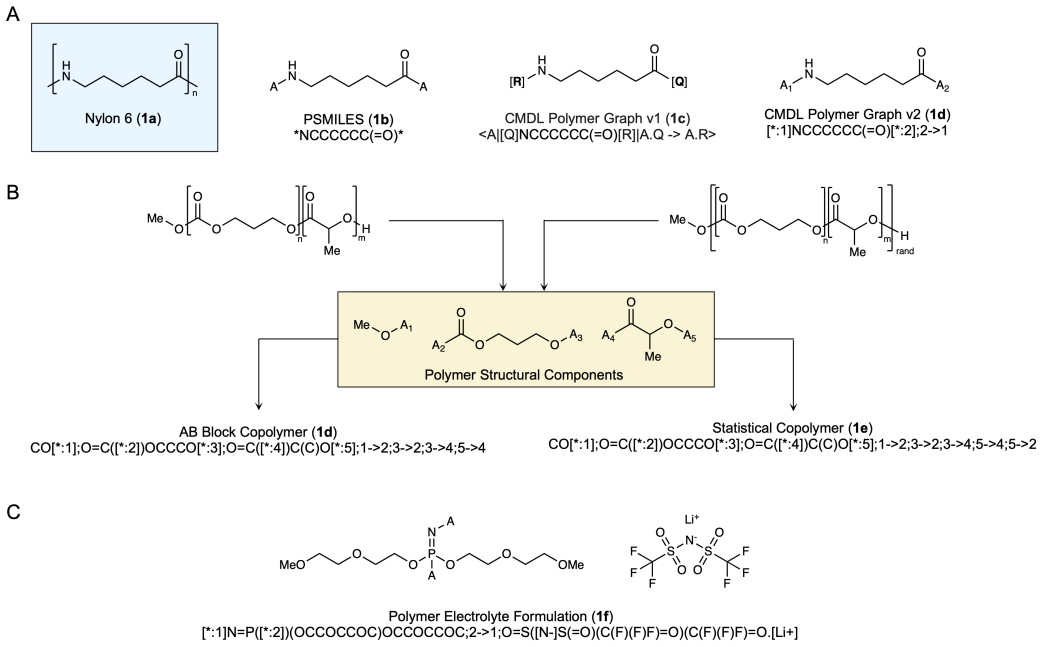

Figure 2: A) Example representations of Nylon 6 (**1a**) as PSMILES (**1b**), the original CMDL polymer graphs (**1c**) and the updated CMDL polymer graphs (**1d**. B) Example polymer graph representation of a poly(trimethylene carbonate-co-lactide) copolymer. C) Example

sources,[1, 2, 6, 11, 12, 13, 14, 15, 17, 21, 29, 32], avoiding whenever possible, datasets containing potentially problematic structures. Next, these PSMILES data were canonicalized using RDKit, converted into the SPG representation, and duplicate entries were removed to afford the final 1M SPG pre-training dataset.

3.3 Model Architecture

We pre-train the SPG-TED$_{289M}$ model using a deep bidirectional transformers-based encoder [33] for polymers token processing, integrated within an encoder-decoder architecture for SPG generation. The hyperparameters of the SPG-TED$_{289M}$ base model are provided in Table 1.

Table 1: SPG-TED$_{289M}$ base architecture specificity.

Hidden size	Attention heads	Layers	Dropout	Normalization
768	12	12	0.2	LayerNorm

Vocab size	# SMILES	# Mol tokens	# Encoder	# Decoder	Total params
2993	91M	4T	47M	242M	289M

To enhance relative encoding, the SPG-TED$_{289M}$ employs a modified version of the RoFormer [34] attention mechanism, where position-dependent rotations R_m are applied to the queries and keys at position m. These rotations are implemented as pointwise multiplications, ensuring minimal computational overhead, as demonstrated in Eq. (1).

$$Attention_m(Q, K, V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle} \tag{1}$$

where Q,K,V are the query, key, and value respectively, and φ is a random feature map.

We start with a sequence of polymers tokens extracted from SPG, each embedded in a 768-dimensional space. The encoder-decoder layer is designed to process molecular token embeddings, represented

as $\mathbf{x} \in \mathbb{R}^{D \times L}$, where D denotes the maximum number of tokens and L represents the embedding space dimension.

In encoder-only models, a mean pooling layer is commonly used to represent tokens as SPG in the latent space. However, this method is constrained by the absence of a natural inversion process for the mean pooling operation. To address this limitation, we propose constructing a latent space representation for SMILES by submersing $\mathbf{x}$ into a latent space, denoted as $\mathbf{z}$, as described in Eq. 2.

$$\mathbf{z} = (\text{LayerNorm}\,(\text{GELU}\,(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)))\,\mathbf{W}_2, \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^L$, $\mathbf{W}_1 \in \mathbb{R}^{D \times L}$, $\mathbf{b}_1 \in \mathbb{R}^L$, $\mathbf{W}_2 \in \mathbb{R}^{L \times L}$, with L denoting the latent space size (specifically, $L = 768$) and D representing the original feature space size (namely, $D = 202$). Subsequently, we can immerse $\mathbf{z}$ back by calculating Eq. 3.

$$\hat{\mathbf{x}} = (\text{LayerNorm}\,(\text{GELU}\,(\mathbf{z}\mathbf{W}_3 + \mathbf{b}_3)))\,\mathbf{W}_4 \tag{3}$$

where $\hat{\mathbf{x}} \in \mathbb{R}^{D \times L}$, $\mathbf{W}_3 \in \mathbb{R}^{L \times L}$, $\mathbf{b}_3 \in \mathbb{R}^L$, $\mathbf{W}_4 \in \mathbb{R}^{L \times D}$.

A language layer (decoder) is employed to process the encoded representation $\hat{\mathbf{x}}$. This layer applies non-linearity and normalization to the input, transforming it into a refined vector. The refined vector is then projected onto a set of logits corresponding to the vocabulary. These logits serve as probabilities for predicting the next token in the molecular sequence, thereby enabling the generation of SPG strings token by token [35]. This approach facilitates the sequential decoding of molecular structures, ensuring that the model captures the underlying polymers syntax effectively.

3.4 Pre-training strategies

Pre-training of the SPG-TED$_{289M}$ model was conducted over 150 epochs using the curated SPG dataset, with a fixed learning rate of 1.6e-4 and a batch size of 256 molecules. The training was distributed across 4 NVIDIA V100 (16G) GPUs, parallelized into 4 nodes using DDP and *torch run*. The process involves two key phases: i) Learning polymer token embeddings through a masking mechanism. ii) Mapping these embeddings into a unified latent space that represents the entire SPG string. This latent space not only captures the structural representation of the SPG but also enables the reconstruction of both individual polymer tokens and the complete SPG strings.

Accordingly, the pre-training process utilizes two distinct loss functions: one associated with the token embeddings, driven by the masking process, and another targeting the encoder-decoder layer, focusing on token reconstruction.

For encoder pre-training we use the masked language model method defined in [33]. Initially 15% of the tokens are selected for possible learning. From that selection, 80% of the tokens are randomly selected and replaced with the [MASK] token, 10% of the tokens are randomly selected to be replaced with a random token, while the remaining 10% of the tokens will be unchanged. The implementation of distinct pre-training strategies has positively impacted the model's efficiency, as demonstrated by the observed improvements in the corresponding loss functions. By optimizing the pre-training phases, we have developed a model that is both robust and highly adept at capturing and reconstructing SPG strings.

4 Experiments

To evaluate the latent space generated by our proposed methodology, we evaluated our proposed foundation model on 28 polymers-specific tasks using a set of 16 datasets from different sources as demonstrated in Table 2. Specifically, we demonstrate the capability of the SPG-TED$_{289M}$ in classification and regression tasks. To ensure an unbiased assessment, we maintained consistency with the original benchmark by adopting identical train/validation/test splits for all tasks.

To fine-tune the SPG-TED$_{289M}$ model, each task was executed using a dedicated NVIDIA V100 GPU with 32 GB of memory. In the next sections, we detail the results obtained during these experiments.

Table 2: Evaluated datasets description

Dataset	Description	Metric	Source
Copolymers (MIT)	DFTB computed electron affinity and ionization potential of copolymers.	RMSE	[15]
IBM-Membrane	Computed thermal and gas permeability properties of polymers.	R^2	[29]
ACS-AMI-Homopolymer-Tg	Tg of homopolymers	RMSE	[36]
Polymer-Refractive-Index	Polymer refractive index	RMSE	[32]
Polymer-Electrolyte-Conductivity (MIT)	Conductivity of polymers and polymer formulations	MAE	[17]
Polymer-Gas-Permeability (NETL)	Gas permeability and selectivity of polymers	MAE	[12]
Polymer-Gas-Permeability (CalTech)	Gas permeability of polymers	R^2	[11]
Polyimide-Tg	Tg of polyimides	MAE	[6]
Polymer-Chain-Bandgap-(Egc)	DFT computed polymer chain bandgap	RMSE	[2, 5]
Polymer-Electron-Affinity-(Eea)	DFT computed electron affinity of polymers	RMSE	[2, 5]
Polymer-Bulk-Bandgap-(Egb)	DFT computed bulk bandgap of polymers	RMSE	[2, 5]
Polymer-Ionization-Energy-(Ei)	DFT computed ionization energy of polymers	RMSE	[2, 5]
Polymer-Dielectric-Constant-(EPS)	DFT computed dielectric constant of polymers	RMSE	[2, 5]
Polymer-Refractive-Index-(Nc)	DFT computed refractive index of polymers	RMSE	[2, 5]
Polymer-Crystallization-Tendency-(Xc)	DFT computed crystallization tendency of polymers	RMSE	[2, 5]
Polymer-Conductivity-(PE-II)	Conductivity of polymers	RMSE	[2]

5 Results and Discussion

In this section, we present a detailed analysis of the results obtained using the SPG-TED$_{289M}$ model across a diverse set of tasks, including copolymer electron affinity and ionization potential, polymer membrane properties, multi-task learning for polymers, refractive index prediction, ionic conductivity, gas permeability, and glass transition temperature. Fig. 3 illustrates a summary of the results obtained during the experiments.

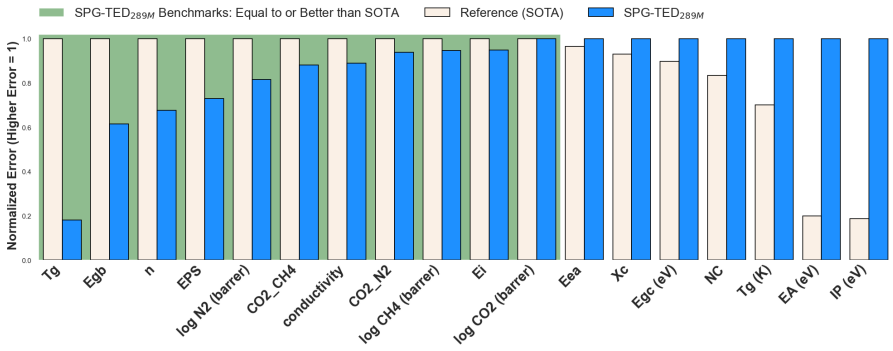

Figure 3: Comparison of the SPG-TED289M model with state-of-the-art models across various polymer property predictions. The results show that SPG-TED289M outperforms SOTA models in **10 out of 18 properties**. The errors are normalized such that a value of 1 represents the maximum error observed in the comparison.

5.1 Comparison with SOTA on benchmarking tasks

Copolymer electron affinity and ionization potential: Here, we present the results of testing the SPG-TED$_{289M}$ model's ability to predict electron affinity (EA) and ionization potential (IP) across a diverse range of monomer compositions, stoichiometries, and chain architectures using a comprehensive copolymer dataset. The dataset encompasses over 40,000 polymers with varying monomer compositions and stoichiometries with density functional tight-binding computed EA and IP, providing a rigorous benchmark for evaluating the model's predictive performance. Table 3 demonstrates the results obtained for this task.

The SPG-TED$_{289M}$ model shows competitive performance, particularly in predicting ionization potential (IP), where it achieves similar accuracy to the Neural Networks (Polymer) model. Although the wD-MPNN slightly outperforms SPG-TED$_{289M}$ in terms of predictive accuracy for both EA and IP, our model still demonstrates solid performance, particularly when considering the challenging nature of the dataset, which includes a wide variety of polymer structures and compositions.

Table 3: Copolymer electron affinity and ionization potential. RMSE is used as evaluation metric, therefore, in this case lower is better. ↓ indicates the best models.

Method	Dataset	
	EA (eV)	IP (eV)
Neural Networks (Monomer) [15]	0.22	0.19
Neural Networks (Polymer) [15]	0.18	0.16
wD-MPNN [15]	**0.03** ↓	**0.03** ↓
SPG-TED$_{289M}$	0.15	0.16

Polymer membranes: For the prediction of polymer membrane properties relevant to carbon dioxide separation, the SPG-TED$_{289M}$ model was fine-tuned to target multiple key properties: half-decomposition temperature ($T_{d\frac{1}{2}}$), glass transition temperature (T_g), and CO_2-permeability ($\log(P_{CO_2})$). These properties are critical for evaluating the performance and stability of polymer membranes in CO_2 separation applications. The results of these fine-tuning experiments are presented in Table 4 demonstrates the results for these complex tasks.

Table 4: Polymer membranes prediction performance. R^2 is used as evaluation metric, therefore, in this case higher values is better. ↑ indicates the best models.

Method	Dataset		
	$T_{d\frac{1}{2}}$	T_g	$\log(P_{CO_2})$
Lasso [29]	0.81	0.90	0.87
ElasticNet [29]	0.81	0.88	0.89
Ridge [29]	0.82	**0.90**↑	**0.90**↑
SPG-TED$_{289M}$ (Frozen Weights)	0.85	0.69	0.71
SPG-TED$_{289M}$ (Fine-tuned)	**0.96**↑	0.86	0.88

As the Table 4 illustrates, the fine-tuned SPG-TED$_{289M}$ model outperforms traditional linear models, such as Lasso, ElasticNet, and Ridge regression, on the regression tasks for these complex molecular properties. Specifically, the fine-tuned model shows significant improvements in prediction accuracy for half-decomposition temperature and CO_2-permeability, demonstrating the effectiveness of the model's architecture in capturing the intricate relationships inherent in polymer data.

The improvement in the results of the fine-tuned model, particularly in the $T_{d\frac{1}{2}}$ predictions, highlights the model's capacity to adapt to the nuances of polymer membrane data, which is crucial for real-world applications in material science.

Polymer multi-task learning: Here, we assess the capabilities of the SPG-TED$_{289M}$ model in multi-task learning, focusing on its ability to predict various polymer properties using high-accuracy density functional theory (DFT) calculations. The model was applied to predict a range of properties, including thermodynamic and physical attributes, optical and dielectric properties, and electronic measurements, as outlined in [5]. Table 5 presents the results of these predictions, where the SPG-TED$_{289M}$ model is compared against the current state-of-the-art (SOTA) methods.

Table 5: Polymer multi-task prediction. RMSE is used as evaluation metric, therefore, in this case lower is better. ↓ indicates the best models.

Method	Dataset						
	Polymer Chain Bandgap (Egc)	Polymer Electron Affinity (Eea)	Polymer Bulk Bandgap (Egb)	Polymer Ionization Energy (Ei)	Polymer Dielectric Constant (EPS)	Polymer Refractive Index (Nc)	Polymer Crystallization Tendency (Xc)
SOTA [2, 3]	**0.44**↓	**0.28**↓	0.49	0.39	0.52	**0.09**↓	**16.57**↓
SPG-TED$_{289M}$	0.49	0.29	**0.32**↓	**0.37**↓	**0.38**↓	0.12	17.82

The results show that the SPG-TED$_{289M}$ model achieves comparable or better accuracy in several key areas, highlighting its effectiveness in predicting complex polymer properties across multiple tasks. This performance demonstrates SPG-TED$_{289M}$ robustness and versatility in capturing the intricate relationships governing polymer properties.

Polymer refractive index: In this experiment, we evaluate the performance of the SPG-TED$_{289M}$ model in predicting the refractive index of polymers, a critical parameter for various optical applications, such as high-refractive-index lenses, which have gained significant interest in recent years. The refractive index is not only important for practical applications but is also theoretically

significant, as it can be determined by the molecule's volume and polarizability, as described by the Lorentz–Lorenz equation. Table 6 presents the results of these predictions, where the SPG-TED$_{289M}$ model is compared against the current state-of-the-art (SOTA) methods.

Table 6: Polymer refractive index prediction. RMSE is used as evaluation metric, therefore, in this case lower is better. ↓ indicates the best models.

Method	Dataset
	Refractive index (n)
GPT-4 [32]	0.0310
Boruta [32]	0.0339
SPG-TED$_{289M}$	**0.0210**↓

Table 6 presents the results of these predictions, where the SPG-TED$_{289M}$ model is compared against current state-of-the-art (SOTA) methods, including GPT-4 and Boruta. The results demonstrate that the SPG-TED$_{289M}$ model outperforms these methods, achieving a lower error rate in refractive index prediction. This improvement suggests that the SPG-TED$_{289M}$ model is highly effective at capturing the underlying relationships between polymer structure and optical properties, making it a valuable tool for designing materials with specific refractive indices.

Polymer ionic conductivity: Here, we evaluate the performance of the SPG-TED$_{289M}$ model in predicting ionic conductivity of solid polymer electrolytes (SPEs). SPEs have the potential to improve lithium-ion batteries by enhancing safety and enabling higher energy densities.[17] However, SPEs suffer from significantly lower ionic conductivity than liquid and solid ceramic electrolytes,limiting their adoption in functional batteries. Machine learning based models can be used To facilitate more rapid discovery of high ionic conductivity SPEs. Table 7 presents the results of these predictions, where the SPG-TED$_{289M}$ model is compared against the current state-of-the-art (SOTA) methods.

Table 7: Polymer ionic conductivity. MAE is used as evaluation metric, therefore, in this case lower is better. ↓ indicates the best models.

Method	Dataset
	Polymer ionic conductivity
XGBoost [32]	1.09
Chemprop [32]	1.08
ChemArr [32]	1.00
SPG-TED$_{289M}$	**0.89**↓

The results highlight the superior predictive capability of the SPG-TED$_{289M}$ model, which achieves a significantly lower error in predicting ionic conductivity. This improvement underscores the model's ability to accurately capture the complex interactions within SPEs that influence their ionic conductivity. The results of SPG-TED$_{289M}$ in this domain not only demonstrates its potential to guide the development of more efficient SPEs, but also supports its broader applicability in the field of battery materials.

Gas permeability of polymers (NETL): In this study, we employed the SPG-TED$_{289M}$ foundation model to evaluate and screen polymers for their effectiveness in CO_2/CH_4 and CO_2/N_2 gas separation using membrane technology. Membrane-based gas separation is a critical process for applications such as carbon capture and natural gas purification, where the selective permeability of polymers can significantly impact efficiency and cost-effectiveness. Table 8 presents the results of these predictions, where the SPG-TED$_{289M}$ model is compared against the current state-of-the-art methods.

Table 8: Gas permeability of polymers (NETL) prediction. MAE is used as evaluation metric, therefore, in this case lower is better. ↓ indicates the best models.

Method	Dataset				
	CO_2	CO_2/CH_4	CH_4	CO_2/N_2	N_2
SOTA [12]	**0.29**↓	5.34	0.37	4.14	0.38
SPG-TED$_{289M}$	**0.29**↓	**4.71**↓	**0.35**↓	**3.89**↓	**0.31**↓

When compared against state-of-the-art (SOTA) methods, SPG-TED$_{289M}$ demonstrates competitive performance, achieving near parity in CO_2 permeability while maintaining strong predictive accuracy

across other metrics. The model's ability to closely approximate SOTA results, particularly in the challenging task of multi-gas separation, underscores its potential utility in the polymer design process for gas separation applications.

Gas permeability of polymers (CalTech): In this study, we assess the SPG-TED$_{289M}$ model's performance in multitask predictions for gas permeabilities of six critical gases: helium (He), hydrogen (H_2), oxygen (O_2), nitrogen (N_2), carbon dioxide (CO_2), and methane (CH_4). Polymer membranes are central to numerous industrial separations, including gas purification and carbon capture, with substantial implications for environmental sustainability. Accurate prediction of gas permeability in polymers is therefore essential for optimizing membrane performance and enhancing the efficiency of these processes.

Table 9: Gas permeability of polymers (CalTech) prediction. R^2 is used as evaluation metric, therefore, in this case higher values is better. ↑ indicates the best models.

Method	Dataset					
	He	H_2	O_2	N_2	CO_2	CH_4
RF (descriptors) [11]	0.73	0.74	0.75	0.74	0.38	0.75
DNN ensemble(descriptors) [11]	0.87	0.88	0.89	0.90	0.90	**0.89**↑
DNN ensemble(MFFs) [11]	0.91	**0.90** ↑	**0.92**↑	0.91	0.90	0.88
SPG-TED$_{289M}$	**0.92**↑	0.87	0.89	**0.91**↑	**0.91**↑	0.85

Results in Table 9, show that the SPG-TED$_{289M}$ model achieves high accuracy in predicting the permeabilities of these gases, closely matching or surpassing the performance of advanced models such as descriptor-based Random Forest (RF) and Deep Neural Network (DNN) ensembles. Notably, SPG-TED$_{289M}$ outperformed existing methods in predicting the permeability of helium and carbon dioxide, which are gases of significant industrial importance due to their roles in energy applications and greenhouse gas mitigation.

Glass-transition temperature of polyimides: In this study, we fine-tuned the SPG-TED$_{289M}$ model to predict the glass transition temperature (T_g) of polymers, a critical property that influences the thermal and mechanical behavior of polymeric materials. Glass transition temperature (T_g) is a fundamental parameter that dictates the thermal performance of polymers, particularly in applications where material stability at various temperatures is crucial. Accurately predicting T_g from the molecular structure of polymer repeating units is essential for the design and development of new polymeric materials with tailored properties for specific industrial applications. The performance of the fine-tuned SPG-TED$_{289M}$ model was benchmarked against current state-of-the-art (SOTA) methods, with the results summarized in Table 10.

Table 10: Glass-transition temperature prediction. MAE is used as evaluation metric, therefore, in this case lower is better. ↓ indicates the best models.

Method	Dataset
	T_g (K)
SOTA [6]	53.02 (24.42)
SPG-TED$_{289M}$	**9.56**↓

The model demonstrated a substantial improvement in prediction accuracy compared to existing SOTA methods, achieving a significantly lower error in Tg predictions. Due to limitations in re-sharing data sourced from PolyInfo, the MAE was computed from separate dataset[16] screened by the author's model whereas the value in the parentheses is the best MAE from the author's model on PolyInfo data.[6] The results of the SPG-TED$_{289M}$ model in this context highlights its ability to learn complex relationships between the polymer structure and its thermophysical properties.

6 Conclusion

This paper presents the development of a SPG-based foundation model for polymers, named SPG-TED$_{289M}$. The model was pre-trained on a curated dataset comprising 1 million SPG samples. In the course of this work, we extended the tokenization process originally introduced by [37] to accommodate the unique characteristics of SPG, resulting in a vocabulary consisting of 2,407 distinct tokens specific to this representation.

The performance of the SPG-TED$_{289M}$ model was comprehensively evaluated across a variety of tasks, including copolymer electron affinity and ionization potential, polymer membrane properties, polymer multi-task learning, refractive index prediction, ionic conductivity, gas permeability, and glass transition temperature. SPG-TED$_{289M}$ model achieved results comparable to specialized models tailored for specific tasks, while requiring only minimal fine-tuning to adapt to the complexities of each task.

References

[1] M. Reis, F. Gusev, N. G. Taylor, S. H. Chung, M. D. Verber, Y. Z. Lee, O. Isayev, and F. A. Leibfarth, "Machine-learning-guided discovery of 19f mri agents enabled by automated copolymer synthesis," *Journal of the American Chemical Society*, vol. 143, no. 42, pp. 17 677–17 689, 2021.

[2] C. Xu, Y. Wang, and A. Barati Farimani, "Transpolymer: a transformer-based language model for polymer property predictions," *npj Computational Materials*, vol. 9, no. 1, p. 64, 2023.

[3] C. Kuenneth and R. Ramprasad, "polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics," *Nature Communications*, vol. 14, no. 1, p. 4099, Jul. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-023-39868-6

[4] F. Wang, W. Guo, M. Cheng, S. Yuan, H. Xu, and Z. Gao, "MMPolymer: A Multimodal Multitask Pretraining Framework for Polymer Property Prediction," Jul. 2024, arXiv:2406.04727 [cond-mat]. [Online]. Available: http://arxiv.org/abs/2406.04727

[5] C. Kuenneth, A. C. Rajan, H. Tran, L. Chen, C. Kim, and R. Ramprasad, "Polymer informatics with multi-task learning," *Patterns*, vol. 2, no. 4, 2021.

[6] Z. Long, H. Lu, and Z. Zhang, "Large-scale glass-transition temperature prediction with an equivariant neural network for screening polymers," *ACS omega*, vol. 9, no. 5, pp. 5452–5462, 2024.

[7] P. Zhang, L. Kearney, D. Bhowmik, Z. Fox, A. K. Naskar, and J. Gounley, "Transferring a Molecular Foundation Model for Polymer Property Predictions," *Journal of Chemical Information and Modeling*, vol. 63, no. 24, pp. 7689–7698, Dec. 2023, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acs.jcim.3c01650

[8] G. Chen, L. Tao, and Y. Li, "Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model," *Polymers*, vol. 13, no. 11, p. 1898, Jan. 2021, number: 11 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2073-4360/13/11/1898

[9] L. Tao, G. Chen, and Y. Li, "Machine learning discovery of high-temperature polymers," *Patterns*, vol. 2, no. 4, p. 100225, Apr. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389921000398

[10] R. Ma, H. Zhang, and T. Luo, "Exploring High Thermal Conductivity Amorphous Polymers Using Reinforcement Learning," *ACS Applied Materials & Interfaces*, vol. 14, no. 13, pp. 15 587–15 598, Apr. 2022, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acsami.1c23610

[11] J. Yang, L. Tao, J. He, J. R. McCutcheon, and Y. Li, "Machine learning enables interpretable discovery of innovative polymers for gas separation membranes," *Science Advances*, vol. 8, no. 29, p. eabn9545, 2022.

[12] S. P. Tiwari, W. Shi, S. Budhathoki, J. Baker, A. K. Sekizkardes, L. Zhu, V. A. Kusuma, D. P. Hopkinson, and J. A. Steckel, "Creation of polymer datasets with targeted backbones for screening of high-performance membranes for gas separation," *Journal of Chemical Information and Modeling*, vol. 64, no. 3, pp. 638–652, 2024.

[13] L. Tao, V. Varshney, and Y. Li, "Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature," *Journal of Chemical Information and Modeling*, vol. 61, no. 11, pp. 5395–5413, Nov. 2021. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c01031

[14] X. Huang, S. Ma, C. Y. Zhao, H. Wang, and S. Ju, "Exploring high thermal conductivity polymers via interpretable machine learning with physical descriptors," *npj Computational Materials*, vol. 9, no. 1, pp. 1–14, Oct. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41524-023-01154-w

[15] M. Aldeghi and C. W. Coley, "A graph representation of molecular ensembles for polymer property prediction," *Chemical Science*, vol. 13, no. 35, pp. 10 486–10 498, 2022.

[16] I. V. Volgin, P. A. Batyr, A. V. Matseevich, A. Y. Dobrovskiy, M. V. Andreeva, V. M. Nazarychev, S. V. Larin, M. Y. Goikhman, Y. V. Vizilter, A. A. Askadskii, and S. V. Lyulin, "Machine Learning with Enormous "Synthetic" Data Sets: Predicting Glass Transition Temperature of Polyimides Using Graph Convolutional Neural Networks," *ACS Omega*, vol. 7, no. 48, pp. 43 678–43 691, Dec. 2022, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acsomega.2c04649

[17] G. Bradford, J. Lopez, J. Ruza, M. A. Stolberg, R. Osterude, J. A. Johnson, R. Gomez-Bombarelli, and Y. Shao-Horn, "Chemistry-informed machine learning for polymer electrolyte discovery," *ACS Central Science*, vol. 9, no. 2, pp. 206–216, 2023.

[18] J. Park, Y. Shim, F. Lee, A. Rammohan, S. Goyal, M. Shim, C. Jeong, and D. S. Kim, "Prediction and Interpretation of Polymer Properties Using the Graph Convolutional Network," *ACS Polymers Au*, vol. 2, no. 4, pp. 213–222, Aug. 2022, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acspolymersau.1c00050

[19] C. Kuenneth, W. Schertzer, and R. Ramprasad, "Copolymer Informatics with Multitask Deep Neural Networks," *Macromolecules*, vol. 54, no. 13, pp. 5957–5961, Jul. 2021, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acs.macromol.1c00728

[20] O. Queen, G. A. McCarver, S. Thatigotla, B. P. Abolins, C. L. Brown, V. Maroulas, and K. D. Vogiatzis, "Polymer graph neural networks for multitask property learning," *npj Computational Materials*, vol. 9, no. 1, pp. 1–10, May 2023, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41524-023-01034-3

[21] S. Lo, M. Seifrid, T. Gaudin, and A. Aspuru-Guzik, "Augmenting Polymer Datasets by Iterative Rearrangement," *Journal of Chemical Information and Modeling*, vol. 63, no. 14, pp. 4266–4276, Jul. 2023, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acs.jcim.3c00144

[22] L. Tao, J. Byrnes, V. Varshney, and Y. Li, "Machine learning strategies for the structure-property relationship of copolymers," *iScience*, vol. 25, no. 7, p. 104585, Jul. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004222008574

[23] T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson, J. A. Kalow, K. F. Jensen, and B. D. Olsen, "BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules," *ACS Central Science*, vol. 5, no. 9, pp. 1523–1531, Sep. 2019, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acscentsci.9b00476

[24] N. H. Park, M. Manica, J. Born, J. L. Hedrick, T. Erdmann, D. Y. Zubarev, N. Adell-Mill, and P. L. Arrechea, "Artificial intelligence driven design of catalysts and materials for ring opening polymerization using a domain-specific language," *Nature Communications*, vol. 14, no. 1, p. 3686, Jun. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-023-39396-3

[25] J. Born and M. Manica, "Regression transformer enables concurrent sequence regression and generation for molecular language modelling," *Nature Machine Intelligence*, pp. 1–13, 2023.

[26] E. Soares, V. Shirasuna, E. V. Brazil, R. Cerqueira, D. Zubarev, and K. Schmidt, "A large encoder-decoder family of foundation models for chemical language," *arXiv preprint arXiv:2407.20267*, 2024.

[27] E. Soares, F. Cipcigan, D. Zubarev, and E. V. Brazil, "A framework for toxic pfas replacement based on gflownet and chemical foundation model," in *NeurIPS 2023 AI for Science Workshop*, 2023.

[28] R. Ma and T. Luo, "PI1M: A Benchmark Database for Polymer Informatics," *Journal of Chemical Information and Modeling*, vol. 60, no. 10, pp. 4684–4690, Oct. 2020. [Online]. Available: https://doi.org/10.1021/acs.jcim.0c00726

[29] R. Giro, H. Hsu, A. Kishimoto, T. Hama, R. F. Neumann, B. Luan, S. Takeda, L. Hamada, and M. B. Steiner, "Ai powered, automated discovery of polymer membranes for carbon capture," *npj Computational Materials*, vol. 9, no. 1, p. 133, 2023.

[30] R. Gurnani, D. Kamal, H. Tran, H. Sahu, K. Scharm, U. Ashraf, and R. Ramprasad, "polyG2G: A Novel Machine Learning Algorithm Applied to the Generative Design of Polymer Dielectrics," *Chemistry of Materials*, vol. 33, no. 17, pp. 7008–7016, Sep. 2021, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acs.chemmater.1c02061

[31] E. Soares, E. V. Brazil, K. F. A. Gutierrez, R. Cerqueira, D. Sanders, K. Schmidt, and D. Zubarev, "Beyond chemical language: A multimodal approach to enhance molecular property prediction," *arXiv preprint arXiv:2306.14919*, 2023.

[32] K. Hatakeyama-Sato, S. Watanabe, N. Yamane, Y. Igarashi, and K. Oyaizu, "Using gpt-4 in parameter selection of polymer informatics: improving predictive accuracy amidst data scarcity and 'ugly duckling'dilemma," *Digital Discovery*, vol. 2, no. 5, pp. 1548–1557, 2023.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[34] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[35] J. Ferrando, G. I. Gállego, I. Tsiamas, and M. R. Costa-jussà, "Explaining how transformers use context to build predictions," *arXiv preprint arXiv:2305.12535*, 2023.

[36] J. Hu, Z. Li, J. Lin, and L. Zhang, "Prediction and interpretability of glass transition temperature of homopolymers by data-augmented graph convolutional neural networks," *ACS Applied Materials & Interfaces*, vol. 15, no. 46, pp. 54 006–54 017, 2023.

[37] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.