https://doi.org/10.1038/s44387-026-00087-1

POLYT5: an encoder-decoder foundation chemical language model for generative polymer design

Harikrishna Sahu, Wei Xiong, Anagha Savit, Shivank S. Shukla & Rampi Ramprasad ✉

Traditional machine learning has advanced polymer discovery, yet direct generation of chemically valid and synthesizable polymers without exhaustive enumeration remains a challenge. Here we present POLYT5, an encoder-decoder chemical language model based on the T5 architecture, trained to understand and generate polymer structures. POLYT5 enables both property prediction and the targeted generation of polymers conditioned on desired property values. We demonstrate its utility for dielectric polymer design, seeking candidates with dielectric constant >3, bandgap >4 eV, and glass transition temperature >400 K, alongside melt-processability and solubility requirements. From over 18,000 generated promising candidates, one was experimentally synthesized and validated, showing strong agreement with predictions. To further enhance usability, we integrated POLYT5 within an agentic AI framework that couples it with a general-purpose LLM, allowing natural language interaction for property prediction and generative design. Together, these advances establish a versatile and accessible framework for accelerated polymer discovery.

Materials discovery is undergoing a transformation with the rise of generative models—particularly large language models (LLMs)—for tasks such as data extraction from scientific literature, property prediction, synthesis planning, and molecular generation[1–6]. A particular class of generative schemes, referred to as "foundation" models, are trained on massive general-purpose language datasets, requiring large vocabularies and architectures with billions of parameters. While powerful, they demand substantial computational resources and lack the domain-specific knowledge essential for materials science. Even when fine-tuned for materials applications, these models can hallucinate, producing syntactically valid but chemically implausible or infeasible outputs[7]. Training on curated datasets encoded with domain-appropriate tokens makes it likely that these models will outperform general-purpose counterparts while using only a fraction of the parameters, reducing computational cost and enhancing reliability, interpretability, and adaptability for downstream materials discovery tasks[8–10].

Advances in generative models for materials discovery have been preceded by the development of the so-called state-of-the-art (SOTA) models. These predictive models have traditionally relied on conventional machine learning (ML) techniques, which typically require carefully engineered numerical fingerprints to represent chemical structure. More recently, polymer-specific encoder language models (e.g., PolyBERT[11] and TransPolymer[12]) trained on large polymer corpora have also been used to generate learned embeddings that serve as effective, transferable fingerprints for downstream prediction tasks. In addition, general-purpose LLMs fine-tuned for specific prediction tasks[3,13] offer advantages over SOTA models by bypassing complex feature engineering through natural language prompts, handling missing features more effectively and rapidly adapting to new problems. In our recent work, we applied this fine-tuning strategy to polymers across multiple property domains, including thermal properties[14], solubility[15], and the performance of organic solar cells[16], demonstrating its potential to streamline predictive modeling in materials informatics.

While accurate property prediction is crucial, the complementary challenge in materials discovery is design, i.e., generating candidate structures that meet desired property or performance criteria. In high-throughput screening powered by ML, candidates—first generated by users based on heuristics—are rapidly evaluated to identify promising materials. However, these approaches still explore only a minute fraction of the vast chemical space and remain constrained by user-imposed biases and limitations inherent in the enumeration process[17–19]. Generative design methods address this limitation by automatically producing structures conditioned on target properties, enabling more efficient navigation of chemical space and accelerating the identification of globally optimal materials. A variety of deep generative modeling approaches have been explored for molecular design, including variational autoencoders (VAEs), generative adversarial networks (GANs), graph neural networks (GNNs), and more recently, transformer-based architectures[6,20]. Dollar et al.[21] showed that incorporating self-attention layers into generative VAE models enables them to learn complex molecular grammar, with individual

School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. ✉e-mail: rampi.ramprasad@mse.gatech.edu

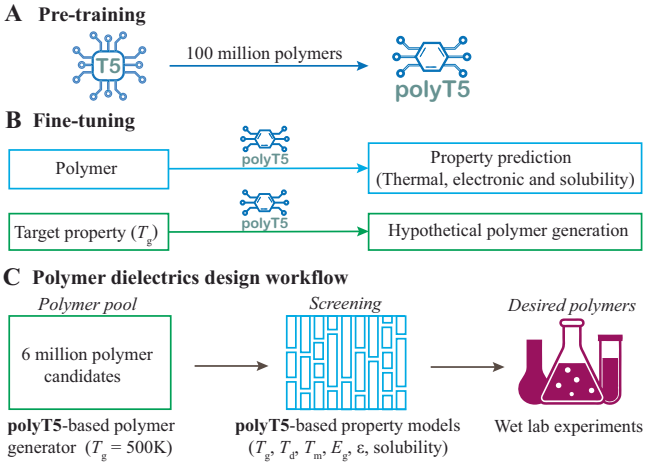

Fig. 1 | Schematic workflow illustrating the large language model (LLM)-based framework for organic material design targeting dielectric applications. **A** POLYT5, a T5-based language model, pre-trained on a corpus of 100 million polymer structures to learn polymer "language" (i.e., chemically valid syntax and recurring structural motifs) via span-masked reconstruction. **B** Fine-tuning POLYT5 for domain-specific tasks, including: (i) property prediction for thermal, electrical, and solubility-related properties and (ii) conditional generation of hypothetical polymer candidates given a target property (e.g., T_g). **C** Application to dielectric polymer discovery: polymers are generated targeting T_g = 500 K and screened using a set of fine-tuned property predictors ("tiles") that jointly enforce dielectric, thermal/processability, and solubility constraints.

attention heads capturing distinct high-level relationships between atomic and structural groups, thereby improving the handling of longer and more complex SMILES (Simplified Molecular Input Line Entry System) strings. He et al.[22] demonstrated the effectiveness of transformer-based models for translating molecules into property-optimized counterparts using SMILES representations. Chemformer[23], a transformer-based model leveraging SMILES representations, demonstrated that self-supervised pre-training and transfer learning can enable efficient fine-tuning across diverse sequence-to-sequence and discriminative cheminformatics tasks, achieving state-of-the-art performance in synthesis prediction and molecular optimization.

For generative molecular design, the choice of string representation for chemical structures is critically important. While SMILES[24,25] has long been the de facto standard for encoding molecules/polymers, its susceptibility to syntactic and semantic errors often leads to invalid structures when used in combination with generative algorithms[26–28]. This limitation can hinder the efficiency and reliability of design workflows. In contrast, SELFIES (SELF-referencIng Embedded Strings)[29] provides a 100% robust alternative that guarantees syntactically and semantically valid chemical structures, irrespective of how the string is modified or generated. This robustness enables the creation of a valid and continuous latent chemical space, making it particularly well-suited for deep generative models. Owing to its robustness, SELFIES is increasingly being adopted in generative molecular design[30–33].

Since SELFIES was originally developed for small molecules, its direct application to polymers was limited. To overcome this, in our recent work we introduced a pseudo-SELFIES notation specifically tailored for polymers, where the termini (*) were substituted with Astatine atoms (At). This molecule-like representation enabled the further training of the SELFIES-TED model[34]—originally designed for molecules but adapted here for polymers—leading to the development of polyBART[35]. The model was then used to design candidate polymers by perturbing the latent space of known structures with targeted properties. However, the approach inherently constrained the generated candidates to remain close to the starting polymer within the learned chemical space.

Despite growing interest in applying deep generative models to materials discovery, critical gaps remain in the domain of complex organic materials—especially polymers. Given their structural diversity and chemical complexity, there is a strong need for a foundation LLM trained specifically on polymers, capable of capturing underlying structural relationships that can be transferred to downstream tasks. From a methodological perspective, many current language models rely on decoder-only architectures, which process sequences unidirectionally and therefore fail to capture the bidirectional chemical dependencies critical for understanding polymer structures. Moreover, common token-level masking strategies are insufficient for learning high-level structural patterns, as they do not capture multi-token dependencies and long-range contextual relationships essential for effective polymer modeling.

The Text-to-Text Transfer Transformer (T5), an encoder–decoder model introduced by Raffel et al.[36], has emerged as a flexible backbone for molecular and materials language modeling because it supports bidirectional encoding and sequence-to-sequence learning with span masking. In polymer informatics, recent T5-based efforts have begun to demonstrate this promise. PolyNC introduced a polymeric prompting framework with a T5 encoder-decoder to enable multitask polymer property prediction and classification[37], while PolyTAO extended this direction toward generative polymer design by learning property-conditioned generation from curated polymer structure-property data[38]. Despite these important advances, a key gap remains: a polymer-native foundation model that is domain-adaptively pre-trained at very large scale on polymer-only structural corpora, enabling robust transfer across diverse downstream polymer prediction and generation tasks.

Here, we present POLYT5, a domain-adapted encoder-decoder language model based on the T5 architecture, trained on over 100 million chemically diverse polymer structures in SELFIES representation of known and hypothetical candidates, spanning 20 functional groups, 8 heteroatoms, with broad representation of aromatic rings and aliphatic chains. Leveraging bidirectional encoder attention, POLYT5 captures long-range structural relationships, while its span-masking strategy more effectively learns missing and next-token patterns than conventional token-level masking.

To demonstrate its practical utility, POLYT5 was applied to dielectric polymer design, an important class of materials for advanced electronic and energy applications. The base model was fine-tuned for two key tasks: (i) hypothetical polymer generation conditioned on target glass transition temperature (T_g) and (ii) prediction of thermal, electronic, and solubility properties. This framework generated over 6 million candidates, which were screened to identify polymers with high dielectric constant, wide bandgap, elevated T_g, and favorable melt-processability and solubility. A schematic of this LLM-driven pipeline for dielectric polymer design is shown in Fig. 1. A top-performing candidate was synthesized and experimentally validated, with measured properties matching predictions, establishing an end-to-end, informatics-driven workflow for accelerated generative discovery of next-generation functional polymers. To further broaden accessibility, we also demonstrate integration of POLYT5 within an agentic AI framework that combines it with a general-purpose LLM, enabling natural language interaction for property prediction and generative design.

Results

This section details the development and evaluation of POLYT5, demonstrating how the model learns polymer chemistry from large-scale data and applies this knowledge to predictive and generative tasks. We first describe the construction of the polymer corpus and foundation model, followed by fine-tuning for property prediction and generative design. The impact of pre-training is then examined, and the framework is applied to dielectric polymer discovery, experimental validation, and integration within an agentic AI environment for interactive design and prediction.

POLYT5: foundation model for polymers

To develop a foundation model for polymers capable of understanding intrinsic structure-property relationships, the first step is to curate a sufficiently large and chemically diverse training dataset. To this end, a large representative dataset of homopolymer chemical structures was

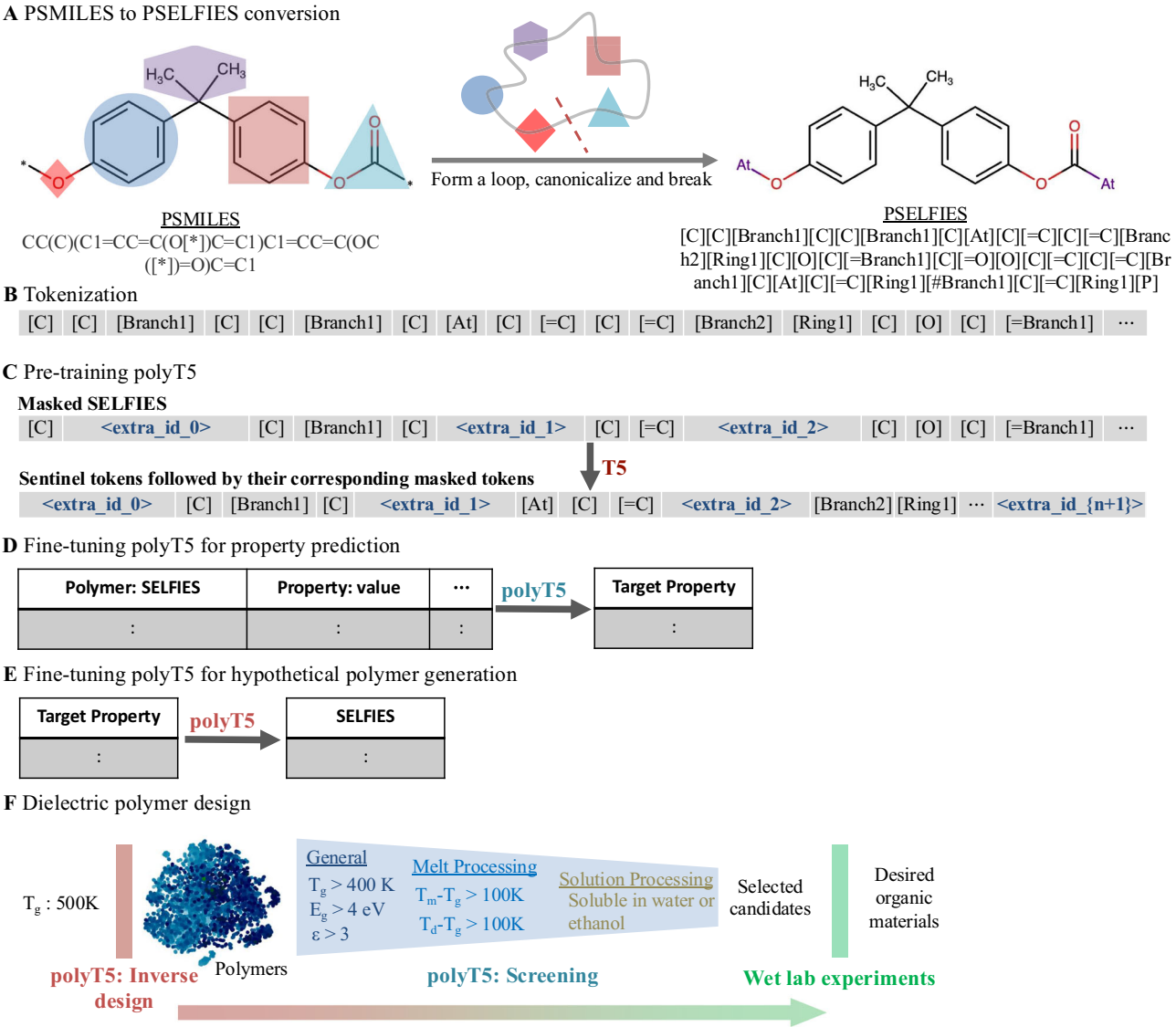

Fig. 2 | POLYT5: key steps in the design of polymers for dielectric applications. **A** Conversion of polymer SMILES (PSMILES) to polymer SELFIES (PSELFIES) representations. **B** Tokenization of PSELFIES for language model input. **C** Span-masked pre-training objective used to train the base POLYT5 model (masked spans replaced by sentinel tokens and reconstructed by the decoder). **D** Fine-tuning of the base model for property prediction tasks, including thermal, electronic, and solubility properties. **E** Fine-tuning of the base model for conditional hypothetical polymer generation (e.g., conditioned on target T_g). **F** Generation and screening of hypothetical polymers targeting desired dielectric performance using a combination of fine-tuned predictors applied as successive filters.

constructed, including 12,473 experimentally synthesized polymers collected from the literature and over 100 million hypothetically generated polymers. To ensure chemical validity and synthetic relevance, the hypothetical polymers were generated by reacting commercially available molecules using well-established reactions such as polycondensation[39–42], click chemistry[43,44], and ring-opening metathesis polymerization (ROMP)[39,45]. These polymers encompass a broad range of functional groups, heteroatoms, and degrees of backbone saturation, as summarized in Table S1.

Each polymer was initially represented using polymer-specific SMILES (PSMILES) notation, where two [*]s denote the terminal ends of the polymer chain. However, since [*] is conventionally used to denote dummy atoms and is not supported in SELFIES—an inherently robust representation well suited for generative modeling of molecules—we developed a custom conversion strategy to overcome this limitation, as illustrated in Fig. 2A[35]. During this process, the two polymer ends were first joined by removing the [*] tokens to form a cyclic structure, followed by canonicalization to mitigate any initial sequence bias. Subsequently, a single bond within the backbone was strategically cleaved, and Astatine (**At**) atoms were attached at the cleavage points. **At** was selected because it is rarely encountered in polymer structures and was entirely absent from our training dataset, minimizing the risk of unintended bias. The resulting pseudo-polymer SMILES— molecular representations with **At** atoms marking the polymer termini— were subsequently converted to SELFIES[29], referred to as PSELFIES. This curated dataset served as the foundation for pre-training the base POLYT5 language model. A custom tokenizer vocabulary was defined to support this representation, comprising 458 tokens including SELFIES tokens (see Fig. 2B), special markers, and additional tokens for property-conditional generation, all integrated as predefined tokens to ensure compatibility with the SentencePiece[46] framework.

To explore the effect of model capacity on downstream performance, we pre-trained three variants of the T5 architecture with progressively increasing sizes, referred to as Small, Medium, and Large, as detailed in the Methods section. These models differ in the number of transformer layers, embedding dimensions, and attention heads, with total parameters ranging from approximately 1.4 million to 59 million (see Table S2). Pre-training

was carried out using a span-masking strategy in which segments of SELFIES strings were replaced with sentinel tokens (Fig. 2C). Across all models, the maximum sequence length was fixed at 200 positions, which is sufficient to cover 99.91% of the SELFIES tokens present in the training dataset, as shown in Fig. S1. This setup enabled a systematic assessment of the trade-off between computational efficiency and representational power for polymer informatics applications.

Figure S2 shows the training loss per batch across two epochs for all three POLYT5 model variants. The loss decreases sharply within the first 0.5 epoch and then plateaus, likely due to the limited token vocabulary (199 unique tokens) used to represent over 100 million PSELFIES, as well as the inherent structural similarity among polymers, despite the underlying chemical diversity present in the training dataset. As expected, the POLYT5-small model exhibits higher loss, while the medium and large variants show marginal differences.

Downstream task 1: property prediction

The base POLYT5 models were fine-tuned for a variety of property prediction tasks (Fig. 2D), covering both regression and classification cases, across thermal, electronic, and solubility properties. The fine-tuning datasets used in this work were compiled and curated in prior studies; unless stated otherwise the labels are experimental, and any simulation/DFT-derived datasets are explicitly indicated[11,15,47,48]. For thermal properties, molecular weight/dispersity and measurement conditions are not consistently reported, so labels should be interpreted as literature-reported values under varying conditions. The corresponding property distributions are shown in Fig. S3[11,35]. The glass transition temperature (T_g) dataset includes 5130 polymers, with values ranging from 80 K to 873 K and a median around 394 K. The distribution is slightly right-skewed, with a higher concentration of polymers between 325 K and 488 K, as shown in Fig. S3A. For decomposition temperature (T_d), the dataset contains 4,204 entries, spanning from 291 K to 1167 K. The melting temperature (T_m) dataset, comprising 2151 polymers, displays a slightly narrower range, from 226 K to 860 K, with most values falling between 400 K and 550 K. The electronic bandgap (E_g) dataset, computed using density functional theory (DFT), consists of 4113 polymers, covering values from as low as 0.07 eV to nearly 9.84 eV. The majority of polymers exhibit E_g values clustered between 3.4 eV and 5.6 eV, with a median value of 4.58 eV (Fig. S3D). The dielectric constant (ε) dataset comprises 1569 polymers, with values ranging from 1.68 to 10.40. The distribution is moderately skewed toward higher values, with a median of 3.09 and most polymers falling between 2.55 and 3.97. The dataset includes both computational and experimental values: 382 dielectric constants were calculated using DFT, while the remaining 1187 values were experimentally measured at nine different frequencies, ranging from 60 Hz to 10^{15} Hz. The distribution of ε values at each measurement frequency, along with the DFT-calculated and total combined distributions, is provided in Fig. S3E. For polymer-solvent solubility, the dataset contains 19,245 soluble and 9970 insoluble cases, covering 6246 unique polymers tested across 58 different solvents.

All three POLYT5 base model variants were fine-tuned to predict thermal (T_g, T_d, T_m), electronic (E_g, ε), and solubility properties. For ε, where the dataset includes both experimental and DFT-derived values, a one-hot identifier indicating the label source (experimental vs DFT) was included during fine-tuning, and model performance is evaluated against experimental targets (with DFT values used as auxiliary supervision). The mean absolute error (MAE) values across epochs for the POLYT5-medium model, evaluated on thermal and electronic properties, are presented in Fig. S4A–C. As expected, the MAE decreased with increasing epochs, saturating around 30 epochs, and also showed a general decrease with larger training set sizes. For the POLYT5-medium model, learning curves showing the RMSE values on unseen test sets are presented in Fig. 3A, while the corresponding plots for R^2 and Pearson correlation coefficient (r) are shown in Fig. S4D–E. Error bars represent the standard deviation obtained from five different random train-test splits. As expected, the RMSE values decrease, and both R^2 and r increase with larger training set sizes. The average RMSEs for an 80% training size on the unseen test set were 40.8, 67.1, and 78.6 for T_g, T_m, and T_d respectively. For E_g and ε, the RMSE values were 0.596 and 0.649, respectively. The parity plots for each property for a representative split with an inset of the error distribution are shown in Fig. 3C–G. As shown in the parity plots, the predicted values closely align with the true values for all properties. The corresponding error distributions are centered around zero with relatively small spreads, indicating good model performance. Overall, the predictive accuracy of POLYT5 is comparable to prior LLM-based baselines reported in the literature[35,47]. In particular, GPT-3.5 and Llama-3, fine-tuned on a similar polymer-property dataset using polymer SMILES as input, report $T_g/T_m/T_d$ errors of 47.2/63.8/80.5 K and 39.5/58.2/77.1 K, respectively, while using polyBART[35] learned embeddings as fingerprints for Gaussian Process Regressor (GPR) models yields 39.9/57.7/71.6 K; for E_g, it reports an RMSE of 0.60.

For the solubility of polymers across various solvents, the learning curves for the prediction accuracy of the soluble, insoluble and overall cases are presented in Fig. 3B, with error bars representing the standard deviation between five random splits. As expected, accuracy improves with increasing training set size. For an 80% training size, the accuracies for soluble and insoluble cases reached 0.957 and 0.917, respectively, resulting in an overall accuracy of 0.943. The confusion matrix for a representative split, shown in Fig. 3H, further illustrates this performance, with 95.7% of soluble and 92.6% of insoluble cases correctly classified. For comparison, Agarwal et al.[15] recently developed a GPT-3.5-based polymer-solvent solubility classifier using polymer SMILES and solvent common names as inputs, reporting accuracies of 0.90 for soluble and 0.83 for insoluble cases. As an additional point of reference, Polymer Genome (PG) predictors[49] provide strong descriptor-based baselines across multiple property domains; while POLYT5 may not always surpass such classical models in predictive accuracy, it offers additional functionality, and PG models can be used as a final filter during candidate selection.

The results for POLYT5-small and POLYT5-large are shown in Figs. S5 and S6, respectively, and a comparison across model sizes is summarized in Table S3. Overall, the fine-tuned POLYT5-small models underperformed compared to the medium model, while the POLYT5-large models showed marginal improvements over the medium variant. Considering the balance between predictive accuracy and computational cost, the POLYT5-medium model was selected for subsequent analyses.

Downstream Task 2: generative design

For the design of dielectric materials, the T_g of polymers plays a critical role in determining their thermal and mechanical stability under operating conditions[50]. Given that the corresponding dataset is the largest in our collection and encompasses a broad range of chemical diversity, it was selected for the generative design task. The POLYT5 models were fine-tuned on the T_g dataset, where T_g values were provided as input and the corresponding PSELFIES strings served as output (Fig. 2E). The fine-tuned models were subsequently employed to generate candidate polymers targeting specific T_g values.

During generation, several hyperparameters influence the selection of the next token and, consequently, the resulting PSELFIES. These include the number of fine-tuning epochs, the T5 temperature parameter (which scales the logits to adjust randomness or confidence in predictions), and nucleus sampling (`top_p`), which limits token selection to the smallest possible set whose cumulative probability meets a specified threshold. To systematically assess how these hyperparameters govern the generation of hypothetical polymers for small, medium and large-POLYT5 models, we varied the number of training epochs from 1 to 15, the temperature from 0.1 to 2.0 in increments of 0.1, and used `top_p` values of 0.75 and 0.95, generating 10,000 polymers for each configuration.

To assess the performance and quality of the generated hypothetical polymers, four evaluation metrics were employed. First, SMILES Validity (SV) was determined using RDKit[51] to ensure the chemical validity of the generated structures. Next, Training Set Deduplication (TSD) filtered out any candidates that were already present in the training dataset. The Dataset

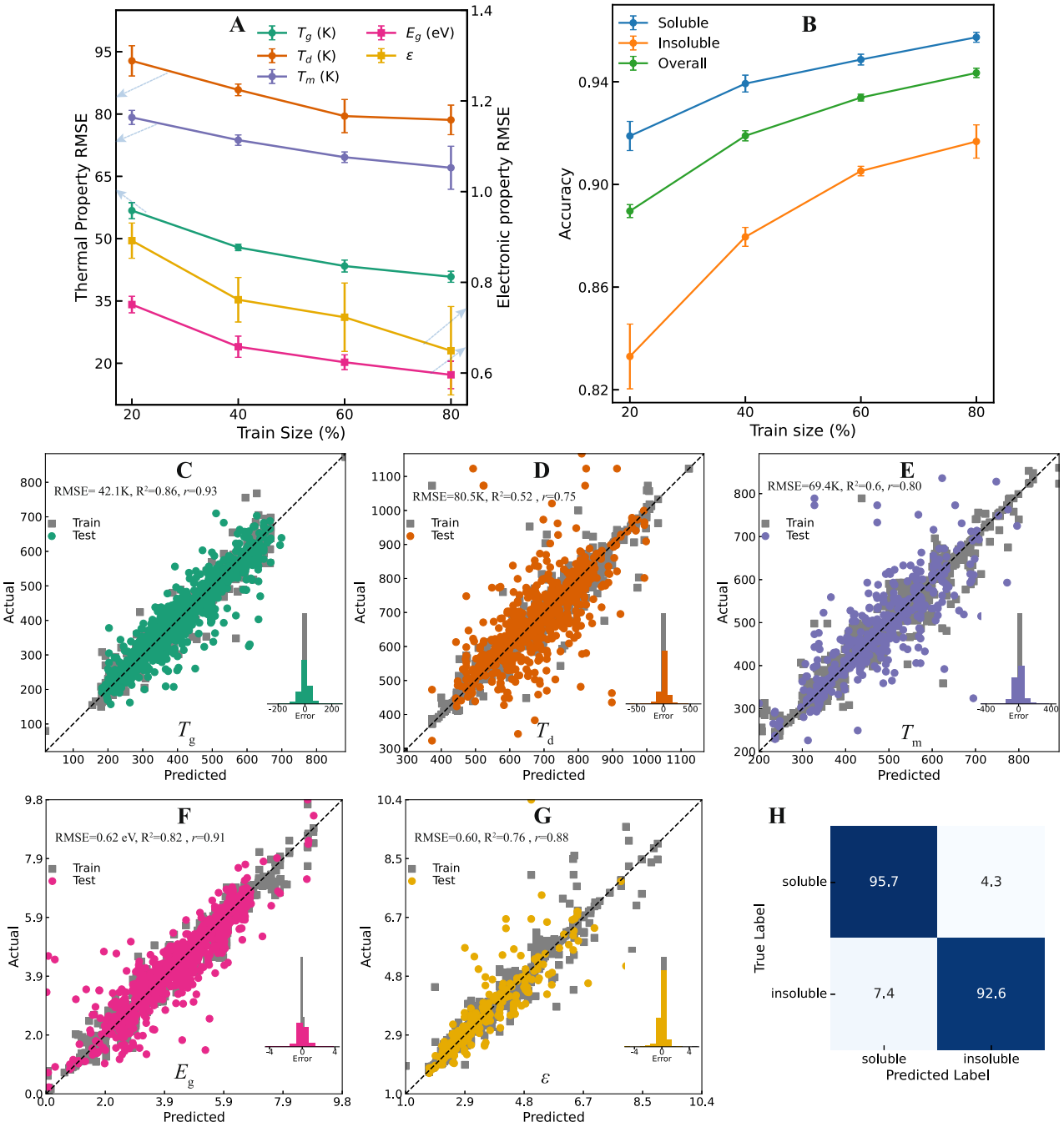

Fig. 3 | Performance of fine-tuned POLYT5-medium models for various property prediction tasks. A Learning curves showing the root-mean-square errors (RMSEs) for thermal and electronic property predictions. **B** Learning curves for solubility prediction. For panels **A** and **B**, error bars represent the standard deviation from five different random train-test splits. **C–G** Parity plots for a representative split, comparing predicted and experimental values for glass transition temperature (T_g), thermal decomposition temperature (T_d), melting temperature (T_m), electronic band gap (E_g), and dielectric constant (ε). (**H**) Confusion matrix for a representative split for solubility prediction, with values reported in percentage.

Deduplication (DD) step removed duplicates within the generated set itself, retaining only unique candidates. Finally, PSMILES Validity (PV) ensured that each retained candidate contained exactly two Astatine (At) atoms, each with a valency of one, as required by the polymer design rules. These filters follow a nested relationship: SV ⊃ TSD ⊃ DD ⊃ PV, such that polymers passing the PV criterion are valid, unique, and novel structures generated by the POLYT5 models.

Figure 4A presents the performance of hypothetical candidate generation by the POLYT5-small, medium, and large models targeting a T_g of 500 K, using the optimal combination of epochs, sampling temperature, and `top_p` values that maximize the number of candidates passing the PV filter. The performance across all possible combinations of these hyperparameters is provided in Fig. S7. Our results revealed important trends in how fine-tuning epoch and generation hyperparameters influence the quality of hypothetical polymer candidates. As the number of fine-tuning epochs increased, the models initially improved, producing fewer invalid candidates. However, beyond a certain point, extended fine-tuning began to increase the number of duplicates within the generated set—a trend especially pronounced for the medium and large models. Higher T5 temperatures increased the likelihood of generating invalid PSMILES, while lower

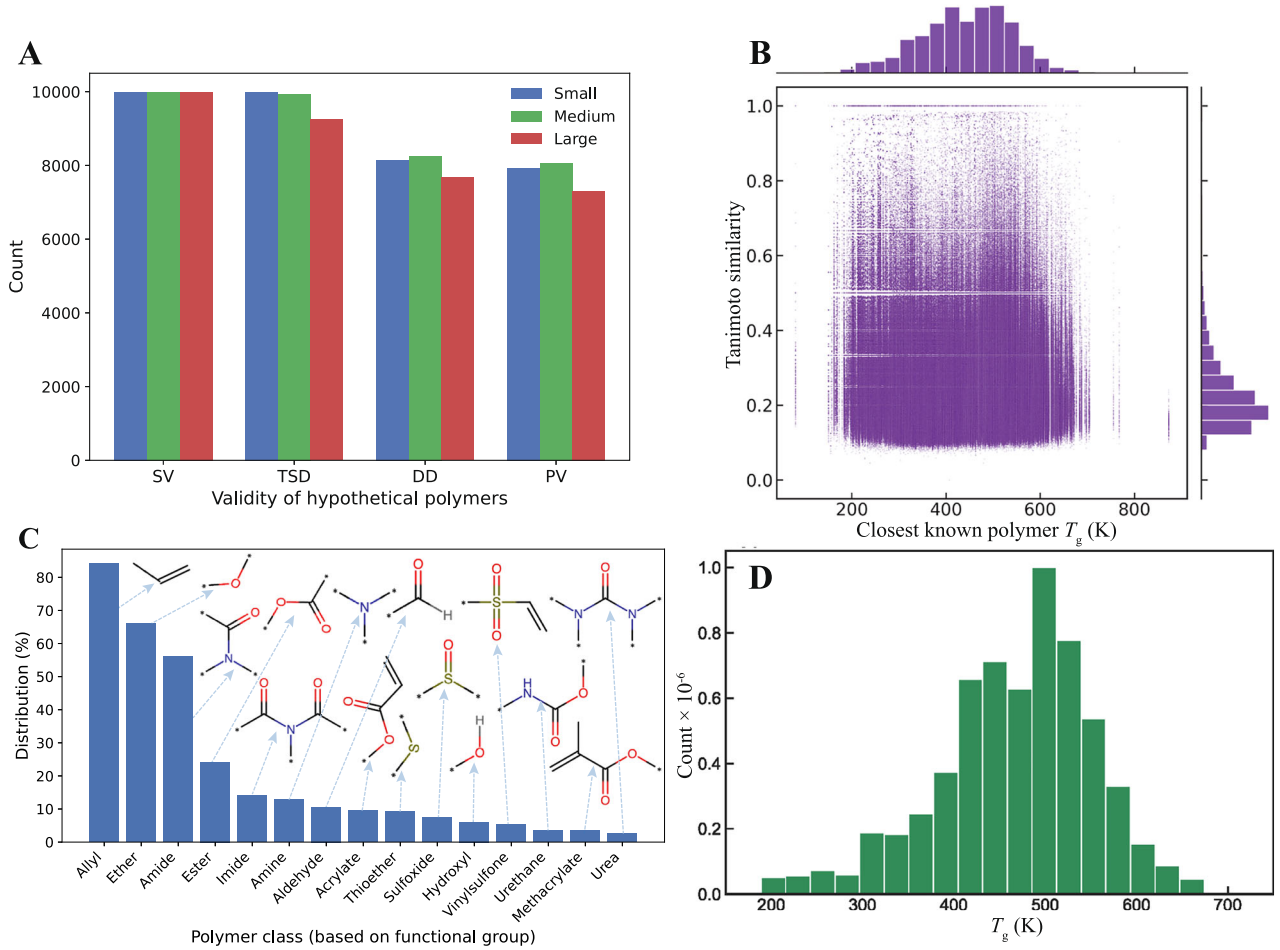

Fig. 4 | Hypothetical candidate generation using POLYT5. A Optimal combination of fine-tuning epoch, T5 sampling temperature, and sampling probability identified for small (14, 0.9, 0.75), medium (6, 1.1, 0.75), and large (8, 1.1, 0.95) POLYT5 models for maximizing the generation of valid hypothetical polymers. SV, TSD, DD, and PV represent successive validation filters: validity of SMILES strings verified by RDKit (SV), removal of duplicates against the training set (TSD), deduplication within the generated hypothetical polymer dataset (DD), and polymer validity (PV) ensuring the presence of exactly two Astatine (At) atoms each with valency one. Note that PV ⊂ DD ⊂ TSD ⊂ SV. **B** Distribution of Tanimoto similarity values between each of the over 6 million POLYT5-generated hypothetical polymers and their closest known polymer in the training dataset, as a function of the closest polymer's T_g (K). **C** Distribution of polymer classes based on the presence of functional groups in the 6 million generated hypothetical polymers. **D** Distribution of predicted T_g values by POLYT5-medium for over 6 million candidate polymers.

temperatures led to excessive duplication, either producing identical hypothetical polymers or reproducing structures from the training set. The effect of `top_p` was also notable: a higher value (0.95) allowed a broader range of tokens during sampling, resulting in greater chemical diversity but also a higher fraction of invalid structures. In contrast, a lower `top_p` (0.75) restricted token choices, reducing invalid outputs but increasing duplication. These observations highlight the delicate trade-offs involved in tuning candidate generation and emphasize the need for a careful balance between model training and generation parameters. Overall, the best balance we observed for POLYT5-medium was obtained at 6 fine-tuning epochs with `top_p`=0.75 and a sampling temperature of T = 1.1, yielding ~ 80.6% of generated candidates passing the PV filter. As a relevant baseline, polyBART[35], which generates candidates by noising the latent space of known polymers rather than conditional generation, reported 86.7% of candidates passing analogous validity/deduplication filters.

In the process of varying fine-tuning epochs and generation hyperparameters, a total of 6,171,066 valid candidate polymers were generated that passed the PV filter. To assess novelty and structural similarity relative to the training set, we computed the Tanimoto similarity between each generated polymer and all polymers in the training data using RDKit-implemented ECFP6 2048-bit fingerprints[51]. To eliminate terminal effects and better approximate infinite polymer chains, the two ends of each repeating unit were connected to form a loop prior to fingerprint calculation. Figure 4B presents the T_g of the most similar (highest Tanimoto similarity) known polymer in the training set against the corresponding similarity value. As expected, most generated polymers are close in T_g to 500 K, consistent with the target value specified during generative design. Importantly, the distribution of Tanimoto similarity values indicates that while generated polymers are guided by examples in the training set, their molecular structures are largely distinct. Further insight into the chemical diversity of the generated polymers is provided in Fig. 4C, which shows the distribution of functional groups present in the generated hypothetical candidates, with a detailed breakdown listed in Table S5. As shown, allyl, ether, and amide functional groups dominate the generated set, while a sizable portion also contains ester, imide, amine, aldehyde, acrylate, and thioether groups. With a cautionary note that the generated chemistries may still be biased toward those prevalent in the fine-tuning dataset, the results overall highlight the model's ability to capture underlying structure-property relationships while generating a chemically diverse set of novel polymer candidates tailored to the desired thermal property.

The fine-tuned POLYT5-medium model was employed to predict the T_g values of all 6 million generated hypothetical polymers, with the resulting distribution shown in Fig. 4D. As expected, the distribution is centered around the target value of 500 K, although it exhibits a slight skew toward lower T_g values, likely due to the higher abundance of training samples below 500 K relative to those above (see Fig. S3). To further assess the

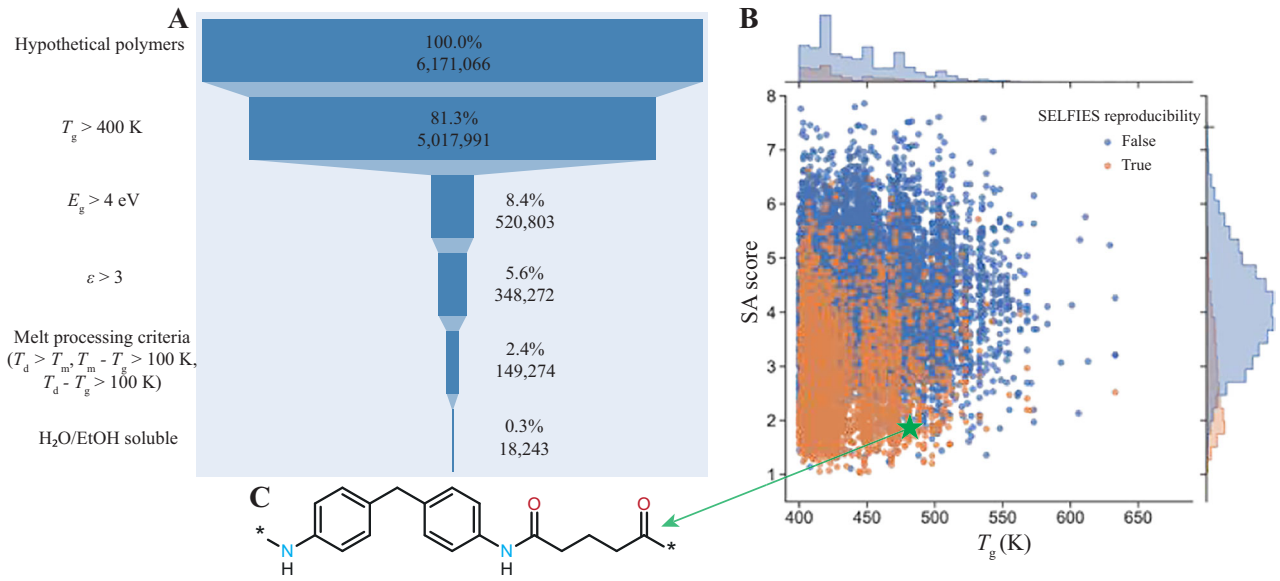

Fig. 5 | Designing dielectric polymers with POLYT5. A Screening of candidates based on thermal, electronic, and solubility criteria. **B** Synthetic accessibility scores of selected polymers versus predicted T_g, annotated by SELFIES reproducibility, with corresponding marginal histograms. **C** Selected polymer for experimental validation.

model's generative capability, an additional set of candidates was generated using POLYT5-medium with a target T_g of 300 K. The predicted T_g distribution for these candidates, shown in Fig. S8, now peaks near 300 K as anticipated. These results collectively demonstrate that POLYT5 has effectively learned the intrinsic relationship between polymer chemical structure and T_g, enabling the generation of candidates tailored to specified target temperatures.

Impact of pre-training: Ablation study on POLYT5

An ablation study was conducted using the POLYT5-medium model to evaluate the impact of pre-training on both property prediction and candidate generation tasks. In this context, "pre-training" denotes the domain-adaptive training described in Section 2.1, performed on 100 million polymer SELFIES using the T5 span-masking objective with sentinel tokens. Table S6 summarizes prediction results for thermal, electronic, and solubility properties. Across all cases, fine-tuning the pre-trained model significantly outperformed training from randomly initialized weights, despite using the same architecture and training protocol. For example, in the case of T_g, the model trained without pre-training achieved an average RMSE of 89.35,K and R^2 of 0.31 across five data splits, while the pre-trained model reduced the RMSE to 40.82,K and improved R^2 to 0.86. Figures S9 and S10 present the generation results for hypothetical polymers targeting a T_g of 500 K. As shown in Fig. S9A–B, without pre-training the model generates a large fraction of SELFIES strings that decode to invalid SMILES, and this issue becomes more severe as the sampling temperature increases. In addition, Fig. S9G–H indicate that pre-training helps the model learn key polymer-specific constraints-most notably the requirement of exactly two **At** atoms, each with valency one-thereby improving the validity of the generated polymer SMILES. The pre-trained model consistently produced a higher number of valid candidates with the desired property (500 ± 50 K), and demonstrated reduced sensitivity to fine-tuning epochs and T5 temperature. Notably, the SELFIES reproducibility (SR) metric—defined as the fraction of generated SELFIES strings that, when converted to SMILES and back to SELFIES, yield the identical string—showed a 5-fold improvement with pre-training. Since SR measures the model's ability to produce canonical SELFIES representations, it reflects a particularly challenging aspect of chemical language modeling. These findings highlight the critical role of large-scale pre-training on 100 million polymers, which equips POLYT5 with an in-depth understanding of structural features in chemical representations and facilitates effective transfer learning for property prediction and polymer generation.

Dielectric polymer design

In designing dielectric polymers, we considered not only a high ϵ (>3) but also key properties such as a wide E_g (>4 eV) and high T_g (>400 K) to ensure thermal and electrical stability. Practical processing criteria, including melt processability ($T_d > T_m$, T_m-T_g >100 K, T_d-T_g >100 K) and solubility either in water or ethanol, were also enforced to ensure that the selected candidates could be synthesized and processed experimentally. These thresholds collectively balance performance, stability, and manufacturability, ensuring that screened polymers are suitable for high-performance dielectric applications.

To identify promising candidates for high-performance dielectric applications, hypothetical polymers were generated using the POLYT5 model fine-tuned for T_g-conditioned generation (Section "Downstream task 2: generative design") with a target of T_g = 500 K, and the resulting set was subsequently screened. This elevated target was chosen to bias generation toward thermally robust candidates (and to ensure a sufficient margin above the T_g >400 K screening threshold). A multi-step screening strategy was then applied (see Fig. 5A) to the resulting 6,171,066 candidates after SV/TSD/DD/PV post-processing; thus, all polymers entering the screening workflow are valid, unique, and non-overlapping with the training set. First, general property criteria were used to narrow down the set: polymers with a predicted T_g exceeding 400 K accounted for approximately 5 million candidates. Of these, 520,803 exhibited a predicted E_g greater than 4 eV, and 348,272 satisfied the target dielectric constant range. Subsequently, 149,274 candidates met the melt processing requirement ($T_d > T_m$,$T_m - T_g$ >100 K, $T_d - T_g$ >100 K). Finally, solubility criteria were applied to prioritize polymers compatible with common, low-cost, and environmentally friendly solvents. Specifically, water and ethanol were selected due to their status as widely available, green solvents[52] frequently used in industrial and laboratory processing. This final filter identified 18,243 polymers predicted to be soluble in at least one of these solvents. This systematic, multi-property screening workflow enabled the identification of a focused set of candidates that combine desirable dielectric, thermal, and processing attributes, suitable for further experimental validation.

Since chemically valid strings can still correspond to highly strained or synthetically impractical motifs, we used the synthetic accessibility (SA)

Table 1 | Properties of the selected polymer as obtained from POLYT5, experiment, and DFT

Property	T_g (K)	T_m (K)	T_d (K)	E_g (eV)
POLYT5	483	603	643	4.45
Experiment/DFT	411	545	651	4.53

Predictions were made using the POLYT5-medium model. The band gap (E_g) was computed using DFT, while the remaining properties were measured experimentally.

score[53] as an additional screening criterion. The SA score combines fragment contributions with a molecular complexity penalty (e.g., large rings/macrocycles, non-standard ring fusions, stereochemical complexity, and molecule size), thereby implicitly penalizing structurally atypical or highly constrained motifs that are more likely to be synthetically challenging or chemically unreasonable. To obtain an SA score for each generated hypothetical polymer, the placeholder atoms ([At] or [*]) were first replaced with hydrogen atoms, effectively treating each polymer as a monomer, and the SA score was then computed using RDKit[51]. As shown in Fig. S11, nearly all known polymers in the training set exhibited SA scores below 6, with most falling within the 2–3 range. Figure 5B presents the distribution of predicted T_g values against SA scores for the screened candidates, with side histograms and annotations based on the SR metric, effectively testing whether POLYT5 generated SELFIES strings in a chemically consistent and robust manner. Among the 18,243 screened candidates, 3142 were found to pass the SR test. Notably, candidates with reproducible SELFIES strings exhibited lower SA scores on average, while those failing the reproducibility test tended to have higher SA scores, including 475 candidates with values exceeding 6. These findings suggest that polymers with reproducible SELFIES representations are generally less structurally complex, making them easier for the model to learn and, in turn, more synthetically accessible.

Experimental validation

The polymer shown in Fig. 5C was selected for experimental verification based on the multi-objective screening criteria and practical synthesis feasibility (e.g., availability of starting materials and a well-established polymerization route). The polymer was prepared by solution polycondensation using glutaryl dichloride and 4,4'-diaminodiphenylmethane under basic conditions. ^{1}H NMR spectroscopy (Fig. S12) confirmed the expected chemical structure and composition. The measured and predicted properties are summarized in Table 1. The experimental T_g was 411 K, in reasonable agreement with the predicted value of 483 K. The T_m appeared at 545 K, about 58 K lower than predicted, suggesting semi-crystalline domains. A higher T_d was observed at 651 K, 8 K higher than the predicted value. From DFT calculations, the E_g was determined to be 4.53 eV, showing remarkable consistency with the predicted 4.45 eV. All deviations fall within the models' error ranges.

Agentic AI framework: integration of LLM and POLYT5 models

To demonstrate how such a polymer generative capability can have enhanced accessibility, we developed an agentic AI framework that integrates a general-purpose large language model (*gpt-5-nano*[54]) with fine-tuned POLYT5 models under a single conversational interface. This framework enables users to query thermal, electronic, solubility, and dielectric properties of polymers, as well as perform generative design tasks, using natural language. All inputs and outputs are expressed in SMILES format, with internal conversion to and from SELFIES for inference. During generation, invalid candidates are automatically filtered, ensuring that only valid polymer SMILES are returned. A schematic overview (Fig. S13) illustrates how user queries are parsed by the controller LLM, routed to the relevant POLYT5 model, and returned as validated outputs.

The framework unifies multiple models under one platform, removing the need to switch between tools, while natural language interaction broadens accessibility to non-experts. It further ensures robustness through automated input handling, reproducibility via standardized schemas, and extensibility through a modular design that allows seamless integration of new models. Together, this agentic AI framework lowers the barrier to advanced polymer modeling, combining the reasoning capabilities of LLMs with the predictive and generative power of POLYT5.

Discussion

In this work, we introduced POLYT5, the first foundation large language model (LLM) tailored for polymers. Trained on over 100 million polymer structures in SELFIES representation with domain-specific tokens using the T5 architecture, POLYT5 was developed in three variants of varying depth and embedding dimensions and fine-tuned for two key downstream tasks: (a) property prediction across thermal, electronic, and solubility properties, and (b) generative design of polymers with targeted glass-transition temperatures (T_g).

For property prediction, POLYT5 achieved RMSEs of 40.82, 67.07 and 78.59 K for T_g, T_m, and T_d, respectively, and 0.60 eV and 0.65 for E_g and ϵ. For polymer solubility classification, the model reached an overall accuracy of 0.94, with soluble and insoluble cases predicted with accuracies of 0.96 and 0.92, respectively. In the generative design task, POLYT5 successfully produced hypothetical polymers with targeted T_g values of 300 and 500 K, with predicted distributions centered near the specified targets. Tanimoto similarity analysis confirmed that the generated candidates were chemically diverse rather than trivial replicas, demonstrating that the model captures structure-property relationships beyond simple memorization.

To demonstrate the practical utility of the developed framework, we applied POLYT5 to the design of high-energy dielectric polymers. Leveraging its generative and predictive capabilities, over 6 million candidates were generated and subsequently screened with property-prediction models. Using cutoff criteria of $\epsilon \geq 3$, $E_g \geq 4$ eV, $T_g \geq 400$ K, along with melt processability and solubility requirements, more than 18,000 candidates were identified as promising. One representative polymer was synthesized and experimentally validated, showing good agreement between measured and predicted properties, while additional DFT calculations confirmed the predicted band gap; together, these validations further support the reliability of the framework.

Beyond model development and validation, we integrated POLYT5 within an agentic AI framework that combines it with a general-purpose LLM, enabling natural language interaction for property prediction and generative design. This conversational interface lowers technical barriers by handling input validation, format conversion, and model selection automatically, thereby making advanced polymer modeling accessible to both experts and non-experts. Taken together, these contributions highlight the promise of domain-specific foundation models in polymer science, POLYT5 captures polymer structure-property relationships and extends this knowledge to prediction and generative design. With fewer than 7.5 million parameters, it achieves both accuracy and efficiency, establishing a practical, accessible, and extensible foundation for accelerated polymer discovery and future applications to more complex materials systems.

POLYT5 currently builds on polymer-specific SELFIES, which provides a robust, syntax-safe representation but also defines the present scope of the framework. The model can be extended beyond homopolymers by fine-tuning on curated datasets for copolymers (e.g., random/block; excluding gradient, ill-defined repeat units or highly complex architectures) and blends, where additional factors (e.g., composition ratios, architecture/topology, or processing descriptors) can be incorporated via specialized tokens alongside SELFIES. In practice, this extensibility is particularly relevant for application areas such as dielectric polymers (electronics/energy storage), thermally stable coatings and composites, and solvent-processable polymers for membranes and separations; translation to these settings will benefit from context-enriched datasets that encode variables beyond repeat-unit chemistry. In principle, multi-property structure generation can be enabled by introducing multiple property-conditioning tokens and fine-tuning on jointly labeled multi-property datasets, allowing direct multi-objective generation rather than post hoc screening. However, the approach is insufficient for polymer systems that cannot be represented as SELFIES sequences, such as cases dominated by conformational/morphological

variability and highly crosslinked networks. POLYT5 also does not currently support BigSMILES; enabling this would require dedicated tokenization and domain-adaptive pretraining on BigSMILES corpora. Finally, because our PSMILES-to-PSELFIES conversion uses Astatine (**At**) atom as a placeholder to represent the two ends of a polymer chain, POLYT5 is not intended to generate **At**-containing polymers, which also have limited practical relevance given At's extreme rarity and radioactivity.

Methods

Generation of reliable hypothetical polymers for pre-training

Over 100 million hypothetical homopolymers were generated using well-established polymerization reactions, encompassing common polymer chemistries such as polyamides, polyimides, polyesters, polyethers, polyureas, and polyurethanes. These candidates were created by reacting commercially available small molecules sourced from databases including eMolecules, ChEMBL, and ZINC-15 via known polycondensation reactions, following strategies similar to those reported in previous studies[39–42,55–58]. To improve chemical plausibility, we applied RDKit-based screening during generation. Specifically, RDKit substructure searches were used to select compatible reactants for each polymerization reaction, and reactant pairs were further filtered using a Gasteiger charge difference criterion (< 0.001) to account for similar condensation reaction rates. In addition, we imposed constraints to screen for backbiting reactions that could lead to looped molecules rather than polymer chains. All reaction rules were implemented using SMARTS in RDKit, which helped avoid chemically infeasible products through RDKit sanitization and valence checks.

To further enhance structural diversity, the ring-opening metathesis polymerization (ROMP) was employed as demonstrated in prior work[39,45]. Several click reactions were also considered, such as Cu-catalyzed azide-alkyne cycloaddition (CuAAC), strain-promoted azide-alkyne cycloaddition (SPAAC), thiol-ene/yne/bromo coupling, Diels-Alder (furan-maleimide), sulfur fluoride exchange (SuFEx), and oxime-based click reactions[43,44,59–64]. These approaches significantly expanded the chemical and functional space of the generated polymer dataset as list in Table S1.

Tokenizer vocabulary

For tokenizing the PSELFIES strings, each substring enclosed within square brackets (e.g., [C], [O]) was treated as a distinct token[35,65], resulting in a base vocabulary of 199 unique tokens. Several special tokens were also introduced, including start- and end-of-sequence markers, unknown and padding tokens, a whitespace marker, and 100 sentinel tokens for masking during pre-training. To further expand the vocabulary and enable property-conditional generation and prediction, an additional 154 tokens were incorporated. These included property names, numerical digits (0-9), decimal point (.), units, arithmetic and relational operators (+, -, >, < , =, etc.), boolean values, and a set of common polymer-related keywords. This resulted in a final vocabulary size of 458 tokens. To ensure compatibility with the SentencePiece[46] tokenizer framework, all SELFIES tokens and additional custom tokens were included as predefined tokens.

Model pre-training: architectures, span Masking, and optimization

We pre-trained three variants of a polymer-specialized T5-based[36] model: POLYT5-small, POLYT5-medium, and POLYT5-large, using a masked span prediction objective within a sequence-to-sequence framework. Polymer structures were represented using SELFIES strings and tokenized using a custom SentencePiece[46] tokenizer comprising 199 unique tokens corresponding to the SELFIES vocabulary. Each model variant follows the standard T5 architecture but differs in size and complexity (see Table S2): POLYT5-small uses an embedding dimension (`d_model`) of 128 with 3 encoder and decoder layers (~1.44 million parameters); POLYT5-medium uses `d_model` = 256 with 4 layers (~7.46 million parameters); and POLYT5-large uses `d_model` = 512 with 8 layers (~58.98 million parameters). All models employ relative positional encodings with a maximum input length of 200 tokens.

The training objective follows the span corruption strategy introduced in the original T5 model. For each polymer sequence, up to 8 non-overlapping masked spans (each up to 3 tokens long) were randomly selected to mask up to 15% of the input tokens. These spans were replaced with sentinel tokens (`<extra_id_n>`) in the input sequence, and the target sequence was constructed by concatenating the masked spans, each prefixed with its corresponding sentinel token. The sentinel tokens were assigned in increasing numerical order of n and placed such that no two masked spans were adjacent, ensuring at least one unmasked token between them.

The models were trained on ~90 million masked polymer sequences (90% of the dataset), using a batch size of 450 for up to 5 epochs on a single NVIDIA L40S GPU. The remaining 10% was reserved for validation and testing. Training loss was calculated using the token-level cross-entropy objective and optimized with the AdamW optimizer. Batch-level training loss was monitored and logged, with checkpoints saved after each epoch. After training, the model weights were saved for subsequent fine-tuning and inference, while a single, pre-defined tokenizer was used throughout.

Downstream task 1: property prediction

To fine-tune POLYT5 for polymer property prediction, we formulated the task as a sequence-to-sequence problem. The model input consisted of polymer structures encoded as SELFIES strings, with supplementary information provided when relevant. For instance, dielectric constant prediction included frequency information in log scale, and solubility prediction incorporated the SELFIES representation of solvents prefixed appropriately. The target output was either a continuous value for thermal or electronic properties, or a categorical label (e.g., "soluble" or "insoluble") for classification tasks such as solubility.

Fine-tuning was performed separately for each property using data splits ranging from 20% to 80% for training and the remainder for testing. Five different random splits were employed to generate learning curves and assess model robustness. Both input and target sequences were tokenized using a single, pre-defined SentencePiece[46] tokenizer, with sequences truncated or padded to a maximum length of 200 tokens.

Training proceeded for up to 30 epochs with a batch size of 16, using the AdamW optimizer with a learning rate of $3e^{-4}$ and a weight decay of 0.01 for regularization. The loss function was token-level cross-entropy, where padding tokens in the target sequences were replaced by −100 to be ignored during loss calculation. Evaluation was conducted at the end of each epoch, measuring mean absolute error (MAE) between predicted and true property values. For regression tasks, predictions were generated using beam search with a beam width of 4, decoded into floating-point numbers, and filtered to remove any invalid or non-numeric outputs.

Model input-output examples for property prediction tasks:

1. **Thermal and electronic Properties** (T_g, T_m, T_d, E_g):
   INPUT: PSELFIES
   Example: `[C] [C] [C] [C] [Branch1] [C] [At] [C] [At]`
   OUTPUT: Property value, e.g., `236.0`
2. **Dielectric constant** (ε):
   INPUT: Property tag with log(frequency) followed by PSELFIES
   Example:`property 4.1; polymer [C] [C] [Branch1] [C] [At] [C] [At]`
   OUTPUT: Property value, e.g., `3.7`
3. **Polymer solubility:**
   INPUT: Polymer and solvent represented in SELFIES format
   Example:`polymer [C] [C] [Branch1] [C] [At] [C] [At]; solvent [C] [C] [C] [O] [C] [Ring1] [Branch1]`
   OUTPUT: Classification token, e.g., `soluble` or `insoluble`

Downstream task 2: polymer generation

For the polymer generation task, we fine-tuned the pre-trained POLYT5 models to generate complete polymer structures in SELFIES format, conditioned on a target property. In this study, T_g was selected as the target property due to its significance in materials design, broad coverage across diverse chemistries, and the relative completeness of the available data in the training set. This task was formulated as a conditional sequence generation problem, where the model learns to map scalar property values to valid polymer sequences.

The fine-tuning dataset consisted of paired (T_g, PSELFIES) examples, with 90% of the data used for training and 10% for validation. All sequences were tokenized using the same SentencePiece tokenizer employed during pre-training and padded to a maximum sequence length of 200 tokens. No masking was applied during this phase, as the model was trained to generate the full target sequence autoregressively, conditioned on the input property string.

Fine-tuning was performed using the HuggingFace Seq2SeqTrainer API for up to 15 epochs, with a batch size of 16. The training used the AdamW optimizer with a learning rate of $3e^{-4}$ and a weight decay of 0.01. Training, evaluation, and checkpointing were performed at the end of each epoch. The objective was to minimize the token-level cross-entropy loss, ignoring padding tokens in the target sequence using a special -100 label. Generation was enabled during evaluation, with a maximum output length of 200 tokens. Upon completion, the fine-tuned models and their associated tokenizers were saved for downstream inference and sampling tasks.

During inference, polymer sequences were generated using a sampling-based approach instead of beam search (which was used for the property prediction task). To systematically assess the effects of temperature and sampling diversity on generation quality, we evaluated models fine-tuned for 1 to 15 epochs using combinations of `top_p` ∈ {0.75, 0.95} sampling thresholds and temperature values ranging from 0.1 to 2.0 in increments of 0.1. This setup enabled a comprehensive exploration of generation behavior under varying stochastic sampling conditions.

Hypothetical polymer generation based on target T_g:

- *INPUT*: Target property value (e.g., `236.0`)
  OUTPUT: Corresponding polymer in SELFIES format
  Example: `[C] [C] [C] [C] [Branch1] [C] [At] [C] [At]`

Agentic AI framework

An agentic AI framework was constructed by integrating a general-purpose LLM (*gpt-5-nano*[54]) with task-specific polymer models. The LLM acted as the controller agent, implemented via the PydanticAI library, which provides structured tool invocation and schema validation. Each POLYT5 model (thermal predictors, Eg predictor, solubility predictor, and Tg-conditioned generator) was wrapped as a PydanticAI Tool with standardized input and output schemas defined using Pydantic BaseModels.

The supervisor LLM was supplied with routing rules in its system prompt to ensure that user queries were directed to the appropriate tool. All inputs and outputs were expressed as SMILES strings: prediction tasks required polymer SMILES with two [*] termini, while solubility additionally required solvent SMILES without termini, and generation required only a target Tg value. In the backend, SMILES were automatically converted to SELFIES for model inference. For generation, candidate SELFIES were converted back to polymer SMILES, with invalid strings rejected to ensure that only valid polymer SMILES were returned. All scientific predictions and generations were executed exclusively by the POLYT5 models.

A lightweight Streamlit interface exposed the framework as a conversational chatbot. User queries were relayed to the supervisor LLM, which invoked the relevant tool and returned outputs displayed as text. The interface retained chat history within each session, allowing the dialog to proceed across successive queries.

First-principles calculations

The initial infinite-chain polymer structure was generated using the PSP[66] package, with vacuum regions of 12 Å along the non-periodic directions to minimize interchain interactions. Density functional theory (DFT) calculations were then performed using the Vienna Ab initio Simulation Package (VASP)[67], employing the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional[68] and a plane-wave energy cutoff of 400 eV. Geometry optimization was considered converged when the maximum force on any atom was below 0.01 eV/Å. The optimized geometries were subsequently used to compute the electronic structure using the HSE06 hybrid functional[69].

Experimental validation

Synthesis of PA from Glutaryl dichloride and 4,4'-Diaminodiphenylmethane. All glasswares were oven-dried and cooled under a nitrogen atmosphere before use. A three-neck round-bottom flask was fitted with a dropping funnel, condenser, and magnetic stir bar, and nitrogen flow was maintained throughout the reaction. Glutaryl dichloride (1.194 g, 7.07 mmol, 1.0 equiv) and 4,4'-diaminodiphenylmethane (1.400 g, 7.07 mmol, 1.0 equiv) were dissolved together in 200 mL of anhydrous dichloromethane (DCM) and cooled in an ice-water bath to 0–5 °C. Triethylamine (0.785 g, 7.78 mmol, 1.1 equiv) was added dropwise over about 30 min under vigorous stirring, keeping the temperature below 10 °C to control the heat release. After TEA addition, the mixture was allowed to reach room temperature (~25 °C) and stirred for an additional 2 h. A white precipitate appeared, indicating polymer formation. The crude product was collected by filtration, washed several times with cold methanol/diethyl ether (1:1 v/v) to remove TEA hydrochloride and unreacted monomers, and then dried in a vacuum oven at 40 °C for 12 h to give the purified PA.

Characterization. The chemical structure and composition of the polymers were confirmed by ^{1}H nuclear magnetic resonance (^{1}H NMR) spectroscopy in d6-DMSO at room temperature. ^{1}H NMR spectra were recorded on a Bruker Avance 500 MHz instrument and calibrated using residual solvent peaks (d6-DMSO: 2.50 ppm for ^{1}H) Characteristic peaks corresponding to both monomer units were observed (Fig. S12A). Thermogravimetric analysis (TGA) was performed on a Pyris 1 TGA under nitrogen at a heating rate of 10 °C/min; the decomposition temperature T_d was defined as $T_{5\%}$ (temperature at 5% mass loss) (Fig. S12B). Differential scanning calorimetry (DSC) was performed on a DSC 3+ STARe system under nitrogen at 10 °C/min using a heat–cool–heat protocol; T_g was taken as the midpoint of the step change in heat flow, and T_m were taken as peak temperatures. Unless stated otherwise, reported DSC transition temperatures correspond to the second heating cycle to minimize thermal-history effects (Fig. S12C–D).

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to IP protection being considered at authors' institution.

Received: 2 November 2025; Accepted: 20 February 2026;
Published online: 03 March 2026

References

1. Zhang, Y. et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artif. Intell.* **1**, 14 (2025).
2. Peivaste, I. et al. Artificial intelligence in materials science and engineering: Current landscape, key challenges, and future trajectories. *Compos. Struct.* **372**, 119419 (2025).
3. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **16**, 2514–2572 (2025).
4. Pyzer-Knapp, E. O. et al. Foundation models for materials discovery –current state and future directions. *npj Comput. Mater.* **11**, 61 (2025).

5. Van, M.-H., Verma, P., Zhao, C. & Wu, X. A survey of AI for materials science: Foundation models, llm agents, datasets, and tools https://arxiv.org/abs/2506.20743 (2025).
6. Anstine, D. M. & Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **145**, 8736–8750 (2023).
7. Zaki, M. & Krishnan, N. A. Mascqa: investigating materials science knowledge of large language models. *Digit. Discov.* **3**, 313–327 (2024).
8. Zhang, J., Chen, X., Ye, X., Yang, Y. & Ai, B. Large language model in materials science: Roles, challenges, and strategic outlook. *Adv. Intell. Discov.* 202500085 (2026).
9. Liu, Z. et al. Molxpt: Wrapping molecules with text for generative pre-training https://arxiv.org/abs/2305.10688 (2023).
10. Chen, H. et al. An overview of domain-specific foundation model: key technologies, applications and challenges https://arxiv.org/abs/2409.04267 (2025).
11. Kuenneth, C. & Ramprasad, R. Polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **14**, 4099 (2023).
12. Xu, C., Wang, Y. & Barati Farimani, A. Transpolymer: a transformer-based language model for polymer property predictions. *npj Comput. Mater.* **9**, 64 (2023).
13. Anisuzzaman, D. M., Malins, J. G., Friedman, P. A. & Attia, Z. I. Fine-tuning large language models for specialized use cases. *Mayo Clin. Proc.Digit. Health* **3**, 100184 (2025).
14. Gupta, S., Mahmood, A., Shukla, S. & Ramprasad, R. Benchmarking large language models for polymer property predictions https://arxiv.org/abs/2506.02129 (2025).
15. Agarwal, S., Mahmood, A. & Ramprasad, R. Polymer solubility prediction using large language models. *ACS Mater. Lett.* **7**, 2017–2023 (2025).
16. Sahu, H., Mahmood, A., Shafique, L. B. & Ramprasad, R. From corpus to innovation: Advancing organic solar cell design with large language models. *npj Comput. Mater.*https://doi.org/10.1038/s41524-025-01896-9 (2025).
17. Chen, G. et al. Machine-learning-assisted de novo design of organic molecules and polymers: opportunities and challenges. *Polymers* **12** https://www.mdpi.com/2073-4360/12/1/163 (2020).
18. Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
19. Sahu, H. et al. Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *J. Mater. Chem. A* **7**, 17480–17488 (2019).
20. Park, H., Li, Z. & Walsh, A. Has generative artificial intelligence solved inverse materials design? *Matter* **7**, 2355–2367 (2024).
21. Dollar, O., Joshi, N., Beck, D. A. C. & Pfaendtner, J. Attention-based generative models for de novo molecular design. *Chem. Sci.* **12**, 8362–8372 (2021).
22. He, J. et al. Molecular optimization by capturing chemist's intuition using deep neural networks. *J. Cheminform.* **13**, 26 (2021).
23. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn Sci. Technol.* **3**, 015022 (2022).
24. Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
25. Weininger, D., Weininger, A. & Weininger, J. L. Smiles. 2. Algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).
26. Sattari, K., Xie, Y. & Lin, J. Data-driven algorithms for inverse design of polymers. *Soft Matter* **17**, 7607–7622 (2021).
27. Batra, R. et al. Polymers for extreme conditions designed using syntax-directed variational autoencoders. *Chem. Mat.* **32**, 10489–10500 (2020).
28. Skinnider, M. A. Invalid smiles are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* **6**, 437–448 (2024).
29. Krenn, M., Häse, F., Nigam, A. K., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (selfies): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).
30. Krenn, M. et al. SELFIES and the future of molecular string representations. *Patterns* **3**, 100588 (2022).
31. Xu, T., Velzeboer, N. & Maruyama, Y. Chemist-computer interaction: Representation learning for chemical design via refinement of SELFIES vae. In *HCI International 2023—Late Breaking Posters*, (eds Stephanidis, C., Antona, M., Ntoa, S. & Salvendy, G.) 353–361 (Springer Nature Switzerland, Cham, 2024).
32. Piao, S., Choi, J., Seo, S. & Park, S. SELF-EdiT: structure-constrained molecular optimisation using SELFIES editing transformer. *Appl. Intell.* **53**, 25868–25880 (2023).
33. Albrijawi, M. T. & Alhajj, R. Lstm-driven drug design using SELFIES for target-focused de novo generation of HIV-1 protease inhibitor candidates for aids treatment. *PLoS One* **19**, 1–30 (2024).
34. IBM Research. https://huggingface.co/ibm-research/materials.selfies-ted.
35. Savit, A., Sahu, H., Shukla, S. S., Xiong, W. & Ramprasad, R. PolyBART: a chemical linguist for polymer property prediction and generative design. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, (Christodoulopoulos, C., Chakraborty, T., Rose, C. & Peng, V.) 12104–12119 (Association for Computational Linguistics, Suzhou, China, 2025).
36. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer https://arxiv.org/abs/1910.10683 (2023).
37. Qiu, H. et al. Polync: a natural and chemical language model for the prediction of unified polymer properties. *Chem. Sci.* **15**, 534–544 (2024).
38. Qiu, H. & Sun, Z.-Y. On-demand reverse design of polymers with polytao. *npj Comput. Mater.* **10**, 273 (2024).
39. Gurnani, R. et al. AI-assisted discovery of high-temperature dielectrics for energy storage. *Nat. Commun.* **15**, 6107 (2024).
40. Kim, S., Schroeder, C. M. & Jackson, N. E. Open macromolecular genome: Generative design of synthetically accessible polymers. *ACS Polym. Au* **3**, 318–330 (2023).
41. Ohno, M., Hayashi, Y., Zhang, Q., Kaneko, Y. & Yoshida, R. Smipoly: Generation of a synthesizable polymer virtual library using rule-based polymerization reactions. *J. Chem. Inf. Model.* **63**, 5539–5548 (2023).
42. Yue, T., He, J. & Li, Y. Polyuniverse: generation of a large-scale polymer library using rule-based polymerization reactions for polymer informatics. *Digit. Discov.* **3**, 2465–2478 (2024).
43. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click chemistry: diverse chemical function from a few good reactions. *Angew. Chem. Int. Ed.* **40**, 2004–2021 (2001).
44. Geng, Z., Shin, J. J., Xi, Y. & Hawker, C. J. Click chemistry strategies for the accelerated synthesis of functional macromolecules. *J. Polym. Sci.* **59**, 963–1042 (2021).
45. Odian, G. *Ring-Opening Polymerization* Chap. 7, 544–618 (John Wiley & Sons, Ltd, 2004).
46. Kudo, T. & Richardson, J. Sentencepiece: a simple and language independent subword tokenizer and detokenizer for neural text processing https://arxiv.org/abs/1808.06226 (2018).
47. Gupta, S., Mahmood, A., Shukla, S. & Ramprasad, R. Benchmarking large language models for polymer property predictions. *Macromol. Rapid Commun*. e00388 (2025).
48. Polyinfo. https://polymer.nims.go.jp/en/.
49. Doan Tran, H. et al. Machine-learning predictions of polymer properties with polymer genome. *J. Appl. Phys.* **128**, 171104 (2020).
50. Liang, Y. et al. All organic polymer dielectrics for high-temperature energy storage from the classification of heat-resistant insulation grades. *J. Polym. Sci.* **61**, 2777–2795 (2023).

51. Landrum, G. Rdkit: Open-source cheminformatics https://www.rdkit.org (2016).
52. Datta, R., Nistane, J., Sose, A., Sahu, H. & Ramprasad, R. Machine learning for green solvents: assessment, selection and substitution. *Adv. Sci*. e16851 (2025).
53. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
54. OpenAI. Gpt-5-nano: Api documentation. https://platform.openai.com/docs/models/gpt-5-nano (2025).
55. Odian, G. *Step Polymerization* Chap. 2, 39–197 (John Wiley & Sons, Ltd, 2004).
56. Sterling, T. & Irwin, J. J. Zinc 15—ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
57. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2016).
58. eMolecules Inc. emolecules database (n.d.). Retrieved May 31, 2024, from https://www.emolecules.com/.
59. Hoyle, C. E. & Bowman, C. N. Thiol–ene click chemistry. *Angew. Chem. Int. Ed.* **49**, 1540–1573 (2010).
60. Lowe, A. B., Hoyle, C. E. & Bowman, C. N. Thiol-yne click chemistry: a powerful and versatile methodology for materials synthesis. *J. Mater. Chem.* **20**, 4745–4750 (2010).
61. Zhang, Y. et al. Thiol–bromo click polymerization for multifunctional polymers: synthesis, light refraction, aggregation-induced emission and explosive detection. *Polym. Chem.* **6**, 97–105 (2015).
62. Gandini, A. The furan/maleimide diels–alder reaction: A versatile click–unclick tool in macromolecular synthesis. *Prog. Polym. Sci.* **38**, 1–29 (2013).
63. Wang, H. et al. SuFEx-based polysulfonate formation from ethenesulfonyl fluoride–amine adducts. *Angew. Chem. Int. Ed.* **56**, 11203–11208 (2017).
64. Collins, J., Xiao, Z., Espinosa-Gomez, A., Fors, B. P. & Connal, L. A. Extremely rapid and versatile synthesis of high molecular weight step growth polymers via oxime click chemistry. *Polym. Chem.* **7**, 2581–2588 (2016).
65. Priyadarsini, I. et al. Self-bart: a transformer-based molecular representation model using SELFIES https://arxiv.org/abs/2410.12348 (2024).
66. Sahu, H., Shen, K.-H., Montoya, J. H., Tran, H. & Ramprasad, R. Polymer structure predictor (psp): a Python toolkit for predicting atomic-level structural models for a range of polymer geometries. *J. Chem. Theory Comput.* **18**, 2737–2748 (2022).
67. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
68. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
69. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).

Acknowledgements

Computations were performed at Expanse (San Diego Supercomputing Center) through an allocation (DMR080044) from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program. The authors acknowledge financial support by the Office of Naval Research through grants N00014-19-1-2103 and N00014-20-1-2175.

Author contributions

H.S. was the primary architect of the project, including dataset development, base and fine-tuned models, candidate generation and screening workflow, and manuscript preparation. A.S. assisted in developing the base model. S.S. generated the hypothetical candidate dataset using known reactions employed for pre-training the base model. W.X. conducted the experimental validations and contributed to writing the corresponding sections. R.R. conceived the project, provided overall guidance, and supervised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-026-00087-1.

Correspondence and requests for materials should be addressed to Rampi Ramprasad.