Exploring Generative Approaches for Predicting Copolymer Sequences from Reaction Conditions

Guanghui Min*
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
jjm8vr@virginia.edu

Wenxin Xu*
Department of Chemistry
University of Virginia
Charlottesville, VA 22903
bkm6ab@virginia.edu

Kateri H. DuBay
Department of Chemistry
University of Virginia
Charlottesville, VA 22903
kateri@virginia.edu

Chen Chen
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
zrh6du@virginia.edu

Abstract

Precise control over monomer sequences in synthetic copolymers is essential for tailoring material properties but remains challenging due to the complexity of polymerization processes. Simulation studies have provided valuable insights into how individual factors influence sequence formation, yet they often examine parameters in isolation and fail to capture their combined effects. Previous applications in polymer sequence design and reaction optimization have proved that machine learning can efficiently navigate complex parameter spaces and accelerate discovery, which is expected to advance the understanding and control of sequence during copolymerization reactions. In this work, we propose a unified conditional block-length distribution generation model to capture the characterization features of polymer sequences, **PolyGen**. Using simulation datasets, we demonstrate that **PolyGen** can accurately predict copolymer block-length distributions in most cases under diverse chemical and physical conditions, including monomer interactions, chain stiffness, activation energy, monomer density, and solvent viscosity. By linking synthesis parameters with sequence outcomes, **PolyGen** establishes a new machine learning–based approach for investigating and guiding the design of sequence-controlled polymers, thereby accelerating their study and potential applications. Our code and dataset are available at https://github.com/GuanghuiMin/PolyGen.

1 Introduction

In polymer science, achieving precise control over the sequences of synthetic copolymers remains a critical yet challenging objective[53, 49, 8]. The composition and monomer sequence along polymer chains jointly determine material properties and, consequently, their potential applications[34, 32]. However, owing to the complexity of synthetic systems, current methods for producing sequence-controlled polymers often suffer from low efficiency and are restricted to specific monomer types[34].

To better understand the complicated factors governing sequence formation during copolymerization, we developed a coarse-grained model to simulate one-pot step-growth copolymerization[60], as de-

*Equal Contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science workshop (NeurIPS 2025).

tailed in the Appendix. Our previous work revealed that relatively weak non-bonded interactions[60], chain stiffness[62], comonomer reactivities[40, 19], and solvent viscosity[19] can influence final sequences by inducing emergent oligomer aggregation, whereas other factors, such as monomer density[19], have negligible effects. Nonetheless, these studies primarily examined only one or two parameters in isolated values, without exploring the combined effects or continuous variation of multiple parameters.

Recent advances in machine learning (ML) have opened new avenues for the investigation and design of sequence-defined polymers[9, 23, 39, 35, 30, 14]. The implementation of effective featurization techniques of polymer sequences, such as through one-hot encoding, realizes the integration of sequence information into ML models, thereby enabling the expansion of data-driven polymer research to sequence-defined copolymers[44]. ML algorithms have been applied to capture complex, nonlinear relationships between monomer sequences and their physicochemical properties[37], enabling the prediction and inverse design of polymer sequences with desired features across vast sequence space[50, 31, 16, 7]. The well-known examples include learning of protein folding landscapes[25, 3], high-throughput peptide drug design[56, 46, 55], and self-assembly material study[57, 4, 45].

Furthermore, ML-driven approaches have shown great promise in learning chemistry reactions and optimizing reaction conditions by identifying subtle dependencies between experimental parameters[64, 38], particularly in the context of organic synthesis[13, 2]. By integrating experimental data, ML tools facilitate efficient exploration of reaction conditions-such as catalysts, solvent, voltage and other controllable factors-which significantly accelerates the identification of optimal conditions for maximizing target product yield. The incorporation of feature importance analysis allows researchers to quantitatively assess the relative impact of individual reaction parameters, offering deeper insights into reaction mechanisms and informing more rational reaction design strategies[2]. As for the study of polymer synthesis, Takasuka *et al.* employed Bayesian optimization to identify process variables that achieve a target monomer composition in radical copolymerization reactions of styrene-methyl methacrylate, which inspires subsequent research applying machine learning to control and elucidate polymerization processes[54].

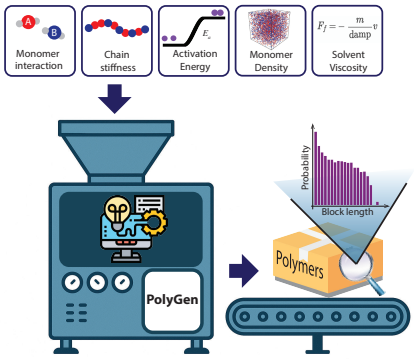


Figure 1: An illustration of the ultimate goal for conditional copolymer chain set prediction. Chemical reaction parameters are fed into a machine learning model, **PolyGen**, which generates the block distribution of possible copolymer chain sets.

However, no studies to date have applied machine learning to examine the interplay of multiple factors influencing sequence formation during copolymerization. This gap hinders the translation of sequences from inverse design ML models into experimentally achievable polymer chains. Without effective understanding about copolymerization reactions, even significant advances in ML-aided sequence design and screening would have limited impact on material discovery and practical application.

In this work, we apply machine learning to investigate the synthesis of sequence-defined polymers using data obtained from our previous coarse-grained simulations of irreversible step-growth copolymerization. We first develop a condition encoder with contrastive learning for the features of resulting copolymer chain sets. This embedding enables quantitative assessment of sequence-set similarity from the block-length distributions. Building on this representation, we then construct a conditional diffusion model that takes simulation parameters as input conditions to generate plausible block-length distributions. This model, named as **PolyGen**, provides a new route for linking polymerization conditions with sequence outcomes, thereby advancing the application of machine learning in sequence-controlled polymer synthesis.

2 Related Works

Polymer Sequence Analysis and Control. Traditional approaches to understanding polymer sequence formation rely heavily on analytical models such as the Mayo-Lewis equation, which describes copolymerization kinetics through reactivity ratios [36, 42]. However, these models assume idealized conditions and provide limited predictive power under complex synthesis environments involving multiple interacting factors [41]. Recent computational studies have revealed that factors beyond simple reactivity ratios—including non-bonded interactions, chain stiffness, and solvent effects—significantly influence sequence development [61, 17, 63]. Despite these insights, existing approaches typically examine parameters in isolation and lack frameworks for predicting sequence outcomes under arbitrary synthesis conditions. Machine learning approaches have shown promise for polymer property prediction and inverse design [10, 24], but most focus on final material properties rather than the underlying sequence formation process during synthesis.

Conditional Generation for Chemical Systems. Generative modeling in chemistry has evolved from early VAE and GAN approaches [15, 6] toward more sophisticated conditional generation frameworks. Recent work has demonstrated the effectiveness of diffusion models for molecular design [1], with Graph Diffusion Transformers enabling multi-conditional generation of small molecules [28]. However, these approaches primarily target discrete molecular structures rather than statistical distributions of structural properties. For polymer systems specifically, generative models have been applied to sequence design [58, 51] and morphology prediction [5], but typically assume known sequence inputs rather than predicting sequence characteristics from synthesis conditions. Conditional generation of polymer block distributions represents a fundamentally different challenge: rather than generating discrete structures, the task requires modeling probability distributions over block lengths while preserving distributional constraints and capturing rare but chemically significant events. Our work addresses this gap by developing the first conditional generation framework specifically designed for polymer block-length distribution prediction from synthesis parameters.

3 Problem Definition

In this section, we first introduce the notations that will be used throughout the paper. We then formally define the notions of *polymer sequence*, *block*, and *copolymer chain set*, which serve as the central objects of our study. Finally, we state our target problem, namely, *conditional block distribution forecasting*.

Notations. Let $\mathcal{A}$ denote a finite alphabet of monomers. A polymer sequence is represented as a string $\mathbf{s} = (s_1, \ldots, s_\ell)$ with $s_i \in \mathcal{A}$ and length $\ell \geq 1$. The set of all finite polymer sequences is written as $\mathcal{A}^\star = \bigcup_{\ell \geq 1} \mathcal{A}^\ell$.

Definition 1 (Polymer Sequence) *A* polymer sequence *is any* $\mathbf{s} \in \mathcal{A}^\star$*, e.g.* $\mathbf{s} = \mathrm{AAABAABA}$.

Definition 2 (Block) *Given a polymer sequence* $\mathbf{s} \in \mathcal{A}^\star$*, a* block *is defined as a maximal contiguous subsequence of identical monomers. Formally,* $\mathbf{s} = (s_1, \ldots, s_\ell)$ *can be uniquely partitioned into blocks*

$$\mathbf{s} = \underbrace{s_1 \cdots s_{i_1}}_{\textit{block 1}} \underbrace{s_{i_1+1} \cdots s_{i_2}}_{\textit{block 2}} \cdots \underbrace{s_{i_{k-1}+1} \cdots s_\ell}_{\textit{block } k}, \tag{1}$$

such that $s_j = s_{i_m}$ *within each block and* $s_{i_m} \neq s_{i_m+1}$ *across adjacent blocks. For instance,* $\mathrm{AAABAABA}$ *consists of blocks* AAA*,* B*,* AA*,* B *and* A*, with block lengths* $(3, 1, 2, 1, 1)$.

Definition 3 (Copolymer Chain Set) *A* copolymer chain set *is a finite multiset*

$$\mathcal{C} = \{\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\}, \quad \mathbf{s}^{(i)} \in \mathcal{A}^\star, \tag{2}$$

representing M *polymer sequences produced under given synthesis conditions. For instance,* $\mathcal{C} = \{\mathrm{AAABAABA}, \mathrm{BBB}, \mathrm{AB}\}$.

Block statistics are central to characterizing copolymer microstructures in polymer chemistry. Classical theories such as the Mayo–Lewis equation describe average reactivity ratios and predict the likelihood of observing alternating vs. blocky arrangements [36, 42]. However, these analytical models typically assume idealized kinetics and provide only coarse expectations (e.g. mean block

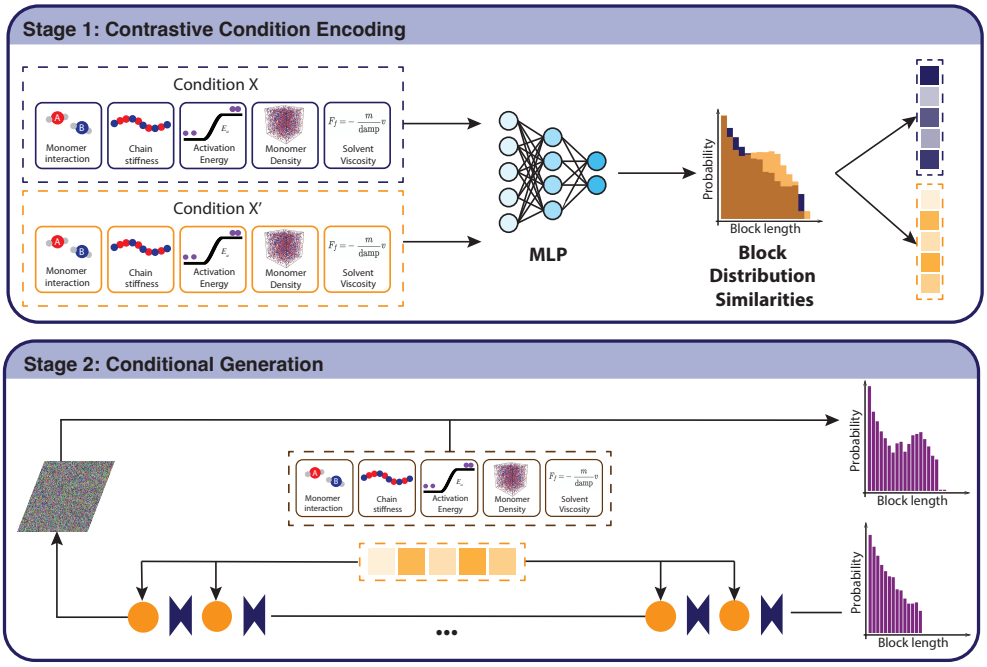


Figure 2: Overall pipeline of our framework. In Stage 1, contrastive condition encoder maps synthesis conditions into a discriminative embedding space, enhancing sensitivity to subtle distributional differences. In Stage 2, a conditional diffusion generator leverages these embeddings to synthesize block-length distributions that align with both the dominant statistics and rare long-tail behaviors.

length). In practice, the full *block length distribution* conveys richer information about microstructure and functionality [20], and serves as a critical descriptor in both simulation and experimental studies. This motivates us to directly forecast block distributions under varying synthesis conditions.

Problem 1 (Conditional Block Distribution Forecasting) *Given synthesis conditions* $\mathbf{X} \in \mathbb{R}^d$ *and an observed copolymer chain set* $\mathcal{C}$*, the goal is to predict the block length distribution associated with* $\mathcal{C}$*:*

$$p_\theta(\mathbf{b} \mid \mathbf{X}), \qquad \mathbf{b} \in \Delta^{K-1}, \tag{3}$$

where $\mathbf{b}$ *is a histogram over block lengths with* K *bins, and* Δ^{K-1} *denotes the probability simplex. At inference time, the model outputs a forecasted block distribution samples from* $p_\theta(\cdot \mid \mathbf{X})$*.*

4 Methodology

Modeling conditional block distributions for copolymer chain sets is challenging due to several factors. First, block length histograms are high-dimensional and sparse, with skewed and long-tailed behavior that standard generative models often fail to capture [27, 29]. Second, rare but chemically meaningful long blocks, though infrequent, are critical to determining material properties [12, 33]. Third, the mapping from synthesis conditions (e.g., monomer ratios, interaction energies) to block statistics is highly nonlinear and governed by polymerization kinetics [36, 42], which makes direct regression approaches insufficient. Finally, supervision is only available at the distribution level (histograms), requiring the model to encode conditions in a way that preserves global statistical structure rather than token-level alignment. To address these challenges, we design a two-stage framework: (i) a contrastive condition encoder that learns semantically meaningful embeddings aligned with block-level statistics, and (ii) a diffusion-based generator that leverages these embeddings to forecast full block distributions under given synthesis conditions.

4.1 Condition Encoder with Contrastive Learning

We learn a condition encoder that maps tabular synthesis conditions $\mathbf{x} \in \mathbb{R}^{d_c}$ to a semantically meaningful representation suitable for conditioning the diffusion model. The embedding should place conditions that induce similar *block-length histograms* close together.

Given $\mathbf{x}$, the encoder applies a multi-layer perceptron (MLP) with LayerNorm and GELU activations to produce a semantic vector

$$\mathbf{h} = f_\phi(\mathbf{x}) \in \mathbb{R}^d, \tag{4}$$

followed by a projection head g_ψ and ℓ_2 normalization for contrastive learning:

$$\mathbf{z} = \frac{g_\psi(\mathbf{h})}{\| g_\psi(\mathbf{h}) \|_2} \in \mathbb{S}^{D-1}. \tag{5}$$

We compare samples by temperature-scaled cosine similarity $s_{ij} = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\tau}$, where a smaller τ yields sharper separation between positives and negatives.

Positive Pair Construction. To ensure the learned embedding reflects chemically meaningful similarities, we define *positive pairs* based on the empirical block-length distributions of copolymer chains. Two condition samples are considered positives if their block histograms are close under the Earth Mover's Distance (EMD), or if they are explicitly paired in the dataset by sharing chains generated under the same conditions. This design grounds the encoder in polymer statistics rather than purely tabular similarity.

Negative Pair Construction. All other in-batch samples and entries from a momentum queue are treated as negatives. In addition, we mine *hard negatives* by selecting those with the smallest EMD to the anchor, sharpening the decision boundary and preventing representation collapse.

Contrastive Objective. The encoder is trained with an InfoNCE-style loss [43] that encourages positive pairs to have higher similarity than negatives. For an anchor condition i with positive set $\mathcal{P}(i)$, the objective is

$$\mathcal{L}_i = -\log \frac{\sum_{j \in \mathcal{P}(i)} \exp(s_{ij})}{\sum_{k \neq i} \exp(s_{ik})}. \tag{6}$$

Minimizing this loss **simultaneously** aligns embeddings of conditions that lead to similar block structures while pushing apart conditions producing dissimilar polymer architectures.

4.2 Diffusion-based Block Distribution Generator

Given encoded synthesis conditions $\mathbf{h}$, we generate realistic block-length distributions $\mathbf{p} \in \Delta^{M-1}$ using a conditional denoising diffusion probabilistic model (DDPM) [21]. To ensure numerical stability, we operate in logit space: $\mathbf{z}_0 = \text{logit}(\mathbf{p} + \epsilon_{\text{smooth}}) \in \mathbb{R}^M$.

Forward Diffusion Process. We progressively corrupt the logits z_0 using a cosine noise schedule over T timesteps:

$$x_t = \sqrt{\alpha_t}\, \mathbf{z}_0 + \sqrt{1 - \alpha_t}\, \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where α_t controls the noise level at timestep t. This produces increasingly noisy representations that gradually erase distributional structure.

Denoising Network. The reverse process employs a one-dimensional diffusion transformer (DiT-1D) [47] with v-parameterization that predicts:

$$v_\theta(x_t, t, \mathbf{h}) \approx \alpha_t \boldsymbol{\epsilon} - \sqrt{1 - \alpha_t}\mathbf{z}_0. \tag{8}$$

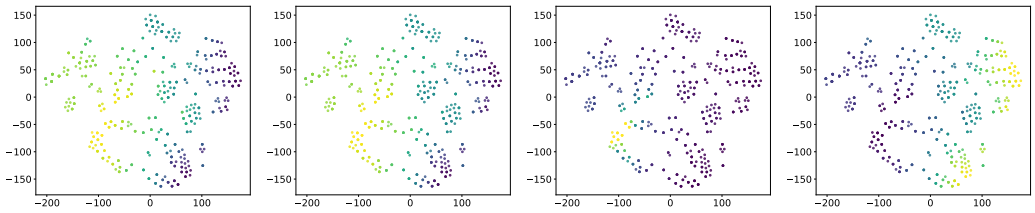
Synthesis condition embeddings $\mathbf{h}$ are injected into each transformer layer via Feature-wise Linear Modulation (FiLM) [48], enabling adaptive generation based on experimental parameters.

Training Objective. Our loss combines denoising accuracy with distributional constraints:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{z}_0, \boldsymbol{\epsilon}} \left[\| v_\theta(x_t, t, \mathbf{h}) - v_{\text{target}} \|_2^2 \right] + \lambda_{\text{EMD}} \mathcal{L}_{\text{EMD}}, \tag{9}$$

where the EMD regularizer preserves distributional shape by aligning cumulative density functions.

Residual Prior. To remove global frequency bias and let the model focus on condition–dependent deviations (e.g., shoulders and long tails), we train the diffusion in a *residual-logit* space. We



(a) p_{AA} (b) p_{BB} (c) Mean block length (d) Alternation index

Figure 3: t-SNE visualization of condition embeddings h learned by our encoder. The embeddings are colored by polymer statistics (normalized): (a) p_{AA}, (b) p_{BB}, (c) mean block length, and (d) alternation index. Darker colors indicate larger values, while lighter colors indicate smaller values.

first estimate a dataset-level prior histogram $\mathbf{p}_{\text{prior}}$ (from the training set) and convert it to logits $z_{\text{prior}} = \text{logit}(\mathbf{p}_{\text{prior}})$. Instead of learning z_0 directly, the target becomes the residual

$$\mathbf{r} \;=\; \mathbf{z}_0 - \mathbf{z}_{\text{prior}}. \tag{10}$$

The DiT then predicts v-parameterized noise on $\mathbf{r}$ via the same objective as above. At inference, we reconstruct logits by adding the prior back,

$$\hat{\mathbf{z}}_0 \;=\; \hat{\mathbf{r}} + \mathbf{z}_{\text{prior}}, \qquad \hat{\mathbf{p}} \;=\; \text{softmax}(\hat{\mathbf{z}}_0/\tau)\,. \tag{11}$$

The residual prior is estimated once from the training set only, ensuring no test-set information leakage.

Classifier-Free Guidance. During training, we randomly drop condition embeddings with probability p_{cfg} to enable classifier-free guidance [22] at inference. At generation time, we use DDIM sampling [52] with guidance scale $w > 1$ to balance sample diversity against conditioning fidelity. The final block distribution is obtained via temperature-scaled softmax: $\hat{\mathbf{p}} = \text{softmax}(\mathbf{z}_0/\tau)$.

5 Empirical Study

5.1 Experimental Settings

Infrastructure and Implementation. All experiments were conducted on a server equipped with NVIDIA A100 GPUs (80GB memory) and an AMD EPYC 7473X CPU with 48 cores and 503GB RAM. Our implementation is based on PyTorch 2.4 with CUDA 12.1.

Dataset. The dataset used to train and evaluate our model was generated from coarse-grained simulations [18, 40, 19] described in Appendix A. It comprises 564 samples, each defined by a set of adjustable parameters summarized in Table 2. All parameters in the dataset are expressed in Lennard–Jones (LJ) reduced units [11, 26], with the length scale set to σ and the energy scale set to ϵ, where 1 ϵ corresponds to $k_{\text{B}}T$ at 100 K.

Evaluation Metrics. To evaluate the accuracy of predicted block-length distributions, we adopt both Kullback–Leibler (KL) divergence and Earth Mover's Distance (EMD). $\text{KL}(\mathbf{p} \,\|\, \hat{\mathbf{p}}) = \sum_{b=1}^{M} p_b \log \frac{p_b}{\hat{p}_b + \varepsilon}$, emphasizes probability calibration by penalizing underestimation of high-density regions. In contrast, $\text{EMD}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{b=1}^{M} |F_{\mathbf{p}}(b) - F_{\hat{\mathbf{p}}}(b)|$, with $F_{\mathbf{p}}$ denoting the cumulative distribution, measures the transport cost of shifting probability mass and is thus sensitive to geometric displacement across bins. Using both metrics provides complementary insights: KL reflects fidelity to true modes, while EMD captures global shape and shift mismatches.

5.2 Performance Analysis

Effectiveness of Contrastive Representation. Figure 3 shows the t-SNE visualization of the learned condition embeddings. Clear clustering structures emerge when colored by different polymer statistics, including p_{AA}, p_{BB}, mean block length, and alternation index. This alignment with chemically meaningful quantities indicates that the encoder successfully captures the underlying polymerization patterns, consistent with the Mayo–Lewis theory [36]. Specifically, the Mayo–Lewis equation

$$\frac{df_A}{dx} = \frac{r_A f_A^2 + f_A f_B}{r_A f_A^2 + 2 f_A f_B + r_B f_B^2} \tag{12}$$

relates the instantaneous copolymer composition to the monomer feed fractions f_A, f_B and the reactivity ratios r_A, r_B. From this, one can derive the transition probabilities $p_{AA} = \frac{r_A f_A}{r_A f_A + f_B}$, $p_{BB} = \frac{r_B f_B}{r_B f_B + f_A}$. Although the previous study from DuBay group proved the failure of Mayo–Lewis theory for the prediction of p_{AA} and p_{BB} with difference in activation energy and non-bonded interactions between like monomer pairs [40], the relation between the pair probabilities and block-level statistics remains true in all of the simulations in the dataset. The probabilities of finding unlike neighbors are therefore calculated from $p_{AB} = 1 - p_{AA}$, $p_{BA} = 1 - p_{BB}$. Furthermore, these transition probabilities influence block-level statistics: the mean block length of A is approximately $\ell_A = 1/(1 - p_{AA})$, the variance is governed by $\mathrm{Var}(\ell_A) = p_{AA}/(1 - p_{AA})^2$, and the alternation index can be summarized as $p_{AB} + p_{BA}$. Since our contrastive learning objective is defined with respect to the block distribution, embeddings that capture these statistics naturally align with chemically meaningful distinctions. In particular, embeddings associated with similar block statistics are mapped closer in the latent space, suggesting that the learned representation is semantically informative and suitable as conditioning input for downstream generative modeling.

Effectiveness of Reconstruction. Figure 4 highlights the reconstruction ability of our diffusion-based generator. In the upper panel, the predicted block-length distribution aligns well with the ground-truth distribution in the high-probability region, indicating that the model successfully captures the dominant statistics. Comparison with the theoretical Mayo–Lewis distribution further shows that the predictions also reproduce non-theoretical features observed in the ground truth, including the long-tail behavior at low probabilities and a characteristic shoulder around 15 monomers corresponding to a preferred block length. The lower panel further supports this observation: most validation samples lie in the lower-left corner of the KL–EMD scatter plot. The KL values are concentrated within the range of 0 to 1, while the EMD mainly fall within the range of 0 to 4, which are small enough to indicate a high degree of agreement between the predicted and ground-truth distributions. These results provide strong evidence that our model successfully captures key distributional features of the block-length histogram, including the long-tail behavior and the shoulder region, which are particularly challenging to model accurately.

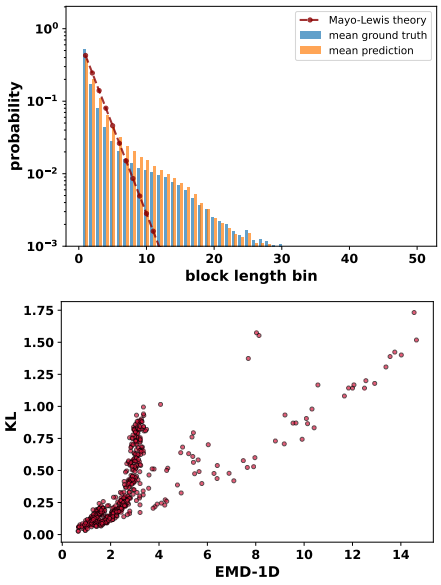


Figure 4: (Top) Mean predicted and ground-truth block-length distributions (log-scale). (Bottom) Scatter plot of validation metrics with KL divergence and 1D EMD.

Effectiveness of Forecasting. Figure 5 presents representative prediction results of our diffusion model. The dataset is randomly divided into a training set (95%) and a held-out test set (5%), where the model is trained only on the training portion and evaluated on unseen test samples. The overlap plots show that in the high-probability regions, the forecasts closely match the ground truth, confirming that the model effectively captures the dominant statistical patterns. The figure also showcases representative prediction results of our model across diverse distributional regimes. For instance, `idx=61` closely aligns with the Mayo–Lewis theoretical distribution, demonstrating that the model can faithfully capture well-characterized behaviors. Other examples highlight the model's ability to forecast challenging distributional patterns beyond the theory: `idx=287` and `idx=320` exhibit pronounced long-tail characteristics, which the model successfully reproduces, while `idx=492` presents a characteristic block length, with the predicted peak magnitude closely matching the ground truth. These results collectively underscore the model's robustness in capturing not only theoretically predictable structures but also complex and rare distributional features that are difficult to model.

5.3 Ablation Study and Sensitivity Analysis

This section aims to assess the effectiveness of the proposed contrastive condition encoding and analyze the sensitivity of critical parameters that shape the accuracy of block distribution prediction.

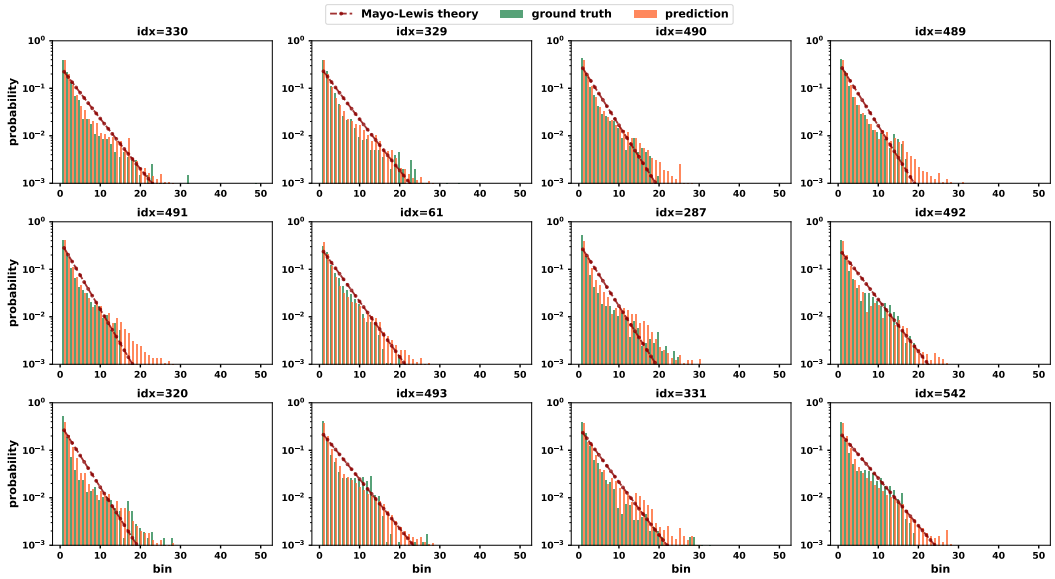


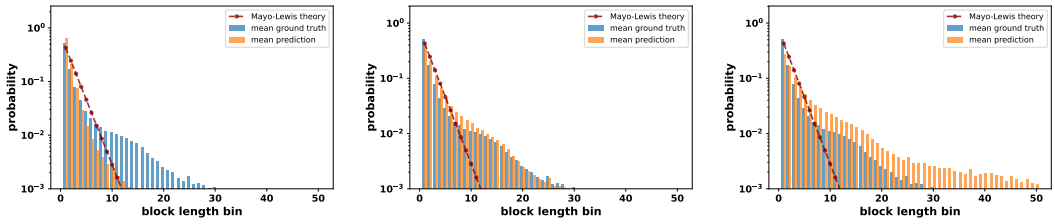
Figure 5: Overlay of predicted and ground-truth block-length distributions for representative test samples.

To verify the effectiveness of contrastive condition encoder, we repeat each experiment three times and compare the differences between the reconstructed block distributions on the training set and the predicted distributions on the test set, with and without the encoding, against the ground truth. Table 1 demonstrates both the effectiveness and necessity of our component.

Table 1: Ablation study of the proposed framework. Both KL divergence and EMD are reported (scaled by 10^{-2}); lower values indicate better performance, and the best results are in **bold**.

Model Variant	KL	EMD
PolyGen		
Reconstruction	**2.89**±0.18	**63.89**±4.16
Prediction	**18.38**±3.17	**95.67**±16.14
`w/o` Cond. Encoder		
Reconstruction	3.02±0.09	65.99±4.17
Prediction	71.52 ±11.01	717.34±316.15

As illustrated in Figure 6, the choice of logit temperature τ has a profound impact on the predicted block-length distribution, particularly in the long-tail regime. Because the final distribution is obtained by $\hat{\mathbf{p}} = \mathrm{softmax}(\mathbf{z}_0/\tau)$, a smaller τ sharpens the softmax and concentrates probability on head bins, thereby underestimating rare long blocks. Conversely, a larger τ increases entropy and spreads mass into the tails, but overly large values may flatten out genuine peaks. We empirically find that setting τ around 1.4 during inference achieves the best balance between head fidelity and tail coverage.



(a) $\tau = 0.8$ (b) $\tau = 1.4$ (c) $\tau = 1.8$

Figure 6: Comparison of mean reconstructed block-length distributions under different logit temperatures. Lower values of τ sharpen the predicted probabilities, while higher values smooth them, influencing how well the model aligns with the ground-truth distribution.

6 Conclusion

This work represents a preliminary step toward developing a generative model that links synthesis conditions with the resulting copolymer sequences. By framing the prediction of block-length distri-

butions as a conditional generation task, we introduce **PolyGen**, a model that integrates contrastive learning for condition encoding and a diffusion-based generator for sequence feature prediction. Our results show that **PolyGen** can capture the dominant statistical patterns of copolymer sequences with satisfying accuracy.

Future efforts will focus on improving the model to enhance sequence feature generation and predictive performance. We also plan to expand and standardize the dataset to include more complex scenarios, such as the incorporation of seed oligomers[59], to enable deeper investigation of sequence control in one-pot synthesis.

Overall, this work proposes a new problem for the application of machine learning in the polymer studies. It also lays the foundation for a machine learning model that not only predicts sequence features but can also be extended to generate full polymer sequences, evaluate factor importance, and support reverse design of synthesis conditions. Such advances will accelerate the rational design and scalable production of sequence-controlled polymers.

References

[1] A. Alakhdar, B. Poczos, and N. Washburn. Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, 64(19):7238–7256, 2024.

[2] A. F. Almeida, F. A. P. Ataíde, R. M. S. Loureiro, R. Moreira, and T. Rodrigues. Augmenting Adaptive Machine Learning with Kinetic Modeling for Reaction Optimization. *The Journal of Organic Chemistry*, 86(20):14192–14198, Oct. 2021.

[3] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. Van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, Aug. 2021.

[4] D. Bhattacharya, D. C. Kleeblatt, A. Statt, and W. F. Reinhart. Predicting aggregate morphology of sequence-defined macromolecules with recurrent neural networks. *Soft Matter*, 18(27):5037–5051, 2022.

[5] D. Bhattacharya, D. C. Kleeblatt, A. Statt, and W. F. Reinhart. Predicting aggregate morphology of sequence-defined macromolecules with recurrent neural networks. *Soft matter*, 18(27):5037–5051, 2022.

[6] N. De Cao and T. Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[7] D. Dellemme, S. Kardas, C. Tonneaux, J. Lernould, M. Fosségré, and M. Surin. From sequence definition to structure-property relationships in discrete synthetic macromolecules: Insights from molecular modeling. 137(23):e202420179.

[8] A. J. DeStefano, R. A. Segalman, and E. C. Davidson. Where Biology and Traditional Polymers Meet: The Potential of Associating Sequence-Defined Polymers for Materials Science. *JACS Au*, 1(10):1556–1571, 2021.

[9] A. L. Ferguson. Machine learning and data science in soft materials engineering. *Journal of Physics: Condensed Matter*, 30(4):043002, Jan. 2018.

[10] A. L. Ferguson. Machine learning and data science in soft materials engineering. *Journal of Physics: Condensed Matter*, 30(4):043002, 2018.

[11] J. Fischer and M. Wendland. On the history of key empirical intermolecular potentials. *Fluid Phase Equilibria*, 573:113876, 2023.

[12] P. J. Flory. *Principles of polymer chemistry*. Cornell university press, 1953.

[13] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Science*, 4(11):1465–1476, Nov. 2018.

[14] W. Ge, R. De Silva, Y. Fan, S. A. Sisson, and M. H. Stenzel. Machine Learning in Polymer Research. *Advanced Materials*, 37(11):2413695, Mar. 2025.

[15] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[16] G. Guzman, J. Ting, S. Murthy, T. Neenan, and S. Kevlahan. ChainSpace: Sequence-aware design for machine learning-driven polymer engineering. *ChemRxiv*. Preprint.

[17] R. L. Hamblin, N. Q. Nguyen, and K. H. DuBay. Selective solvent conditions influence sequence development and supramolecular assembly in step-growth copolymerization. *Soft Matter*, 18(5):943–955, 2022.

[18] R. L. Hamblin, N. Q. Nguyen, and K. H. DuBay. Selective solvent conditions influence sequence development and supramolecular assembly in step-growth copolymerization. *Soft Matter*, 18(5):943–955, 2022.

[19] R. L. Hamblin, Z. Zhang, and K. H. DuBay. Characteristic System Time Scales Can Influence the Collective Sequence Development of Nematically Ordered Copolymers. *Macromolecules*, 57(21):9984–9998, 2024.

[20] I. W. Hamley. *The physics of block copolymers*. Oxford University Press, 1998.

[21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[22] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[23] N. E. Jackson, M. A. Webb, and J. J. De Pablo. Recent advances in machine learning towards multiscale soft materials design. *Current Opinion in Chemical Engineering*, 23:106–114, Mar. 2019.

[24] N. E. Jackson, M. A. Webb, and J. J. De Pablo. Recent advances in machine learning towards multiscale soft materials design. *Current Opinion in Chemical Engineering*, 23:106–114, 2019.

[25] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021.

[26] J. Lenhard, S. Stephan, and H. Hasse. On the history of the lennard-jones potential. *Annalen der Physik*, 536(6):2400115, 2024.

[27] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018.

[28] G. Liu, J. Xu, T. Luo, and M. Jiang. Graph diffusion transformers for multi-conditional molecular generation. *Advances in Neural Information Processing Systems*, 37:8065–8092, 2024.

[29] S. Liu, X. Zhou, Y. Jiao, and J. Huang. Wasserstein generative learning of conditional distribution. *arXiv preprint arXiv:2112.10039*, 2021.

[30] T. Long, Q. Pang, Y. Deng, X. Pang, Y. Zhang, R. Yang, and C. Zhou. Recent Progress of Artificial Intelligence Application in Polymer Materials. *Polymers*, 17(12):1667, June 2025.

[31] S. Luo, X. Jia, and X. Xu. Sequence-based computational design of high-affinity amphiphilic copolymers for protein targeting: A machine learning and coarse-grained simulation approach. 58(14):7005–7016.

[32] J. Lutz. Defining the Field of Sequence-Controlled Polymers. *Macromolecular Rapid Communications*, 38(24):1700582, Dec. 2017.

[33] J.-F. Lutz. Sequence-controlled polymerizations: the next holy grail in polymer science? *Polymer Chemistry*, 1(1):55–62, 2010.

[34] J.-F. Lutz, M. Ouchi, D. R. Liu, and M. Sawamoto. Sequence-Controlled Polymers. *Science*, 341(6146):1238149, 2013.

[35] T. B. Martin and D. J. Audus. Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polymers Au*, 3(3):239–258, June 2023.

[36] F. R. Mayo and F. M. Lewis. Copolymerization. i. a basis for comparing the behavior of monomers in copolymerization; the copolymerization of styrene and methyl methacrylate. *Journal of the American Chemical Society*, 66(9):1594–1601, 1944.

[37] V. Meenakshisundaram, J.-H. Hung, T. K. Patra, and D. S. Simmons. Designing Sequence-Specific Copolymer Compatibilizers Using a Molecular-Dynamics-Simulation-Based Genetic Algorithm. *Macromolecules*, 50(3):1155–1166, Feb. 2017.

[38] M. Meuwly. Machine Learning for Chemical Reactions. *Chemical Reviews*, 121(16):10218–10239, Aug. 2021.

[39] D. Nguyen, L. Tao, and Y. Li. Integration of Machine Learning and Coarse-Grained Molecular Simulations for Polymer Materials: Physical Understandings and Molecular Design. *Frontiers in Chemistry*, 9:820417, Jan. 2022.

[40] N. Q. Nguyen, R. L. Hamblin, and K. H. DuBay. Emergent Sequence Biasing in Step-Growth Copolymerization: Influence of Non-Bonded Interactions and Comonomer Reactivities. *The Journal of Physical Chemistry B*, 126(34):6585–6597, 2022.

[41] N. Q. Nguyen, R. L. Hamblin, and K. H. DuBay. Emergent sequence biasing in step-growth copolymerization: Influence of non-bonded interactions and comonomer reactivities. *The Journal of Physical Chemistry B*, 126(34):6585–6597, 2022.

[42] G. Odian. *Principles of polymerization*. John Wiley & Sons, 2004.

[43] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[44] R. A. Patel, C. H. Borca, and M. A. Webb. Featurization strategies for polymer sequence or composition design by machine learning. *Molecular Systems Design & Engineering*, 7(6):661–676, 2022.

[45] R. A. Patel, S. Colmenares, and M. A. Webb. Sequence Patterning, Morphology, and Dispersity in Single-Chain Nanoparticles: Insights from Simulation and Machine Learning. *ACS Polymers Au*, 3(3):284–294, June 2023.

[46] R. A. Patel and M. A. Webb. Data-Driven Design of Polymer-Based Biomaterials: High-throughput Simulation, Experimentation, and Machine Learning. *ACS Applied Bio Materials*, 7(2):510–527, Feb. 2024.

[47] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[48] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[49] S. L. Perry and C. E. Sing. 100th Anniversary of Macromolecular Science Viewpoint: Opportunities in the Physics of Sequence-Defined Polymers. *ACS Macro Letters*, 9(2):216–225, Feb. 2020.

[50] P. S. Ramesh and T. K. Patra. Polymer sequence design *via* molecular simulation-based active learning. *Soft Matter*, 19(2):282–294, 2023.

[51] P. S. Ramesh and T. K. Patra. Polymer sequence design via molecular simulation-based active learning. *Soft Matter*, 19(2):282–294, 2023.

[52] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[53] J. K. Szymański, Y. M. Abul-Haija, and L. Cronin. Exploring Strategies To Bias Sequence in Natural and Synthetic Oligomers and Polymers. *Accounts of Chemical Research*, 51(3):649–658, Mar. 2018.

[54] S. Takasuka, S. Ito, S. Oikawa, Y. Harashima, T. Takayama, A. Nag, A. Wakiuchi, T. Ando, T. Sugawara, M. Hatanaka, T. Miyao, T. Matsubara, Y.-Y. Ohnishi, H. Ajiro, and M. Fujii. Bayesian optimization of radical polymerization reactions in a flow synthesis system. *Science and Technology of Advanced Materials: Methods*, 4(1):2425178, Dec. 2024.

[55] F. Wan, F. Wong, J. J. Collins, and C. De La Fuente-Nunez. Machine learning for antimicrobial peptide identification and design. *Nature Reviews Bioengineering*, 2(5):392–407, Feb. 2024.

[56] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, Aug. 2023.

[57] M. A. Webb, N. E. Jackson, P. S. Gil, and J. J. De Pablo. Targeted sequence design within the coarse-grained polymer genome. *Science Advances*, 6(43):eabc6216, Oct. 2020.

[58] M. A. Webb, N. E. Jackson, P. S. Gil, and J. J. de Pablo. Targeted sequence design within the coarse-grained polymer genome. *Science advances*, 6(43):eabc6216, 2020.

[59] W. Xu, N. Q. Nguyen, and K. H. DuBay. Seed oligomers regulate sequence development through a templating effect in simulated irreversible step-growth copolymerization. *ChemRxiv*, 2025. Preprint.

[60] Z. Zhang and K. H. DuBay. Modeling the Influence of Emergent and Self-Limiting Phase Separations among Nascent Oligomers on Polymer Sequences Formed during Irreversible Step-Growth Copolymerizations. *Macromolecules*, 52(15):5480–5490, 2019.

[61] Z. Zhang and K. H. DuBay. Modeling the influence of emergent and self-limiting phase separations among nascent oligomers on polymer sequences formed during irreversible step-growth copolymerizations. *Macromolecules*, 52(15):5480–5490, 2019.

[62] Z. Zhang and K. H. DuBay. The Sequence of a Step-Growth Copolymer Can Be Influenced by Its Own Persistence Length. *J. Phys. Chem. B*, 125(13):3426–3437, Apr. 2021.

[63] Z. Zhang and K. H. DuBay. The sequence of a step-growth copolymer can be influenced by its own persistence length. *Journal of Physical Chemistry B*, 125(13):3426–3437, 2021.

[64] Z. Zhou, X. Li, and R. N. Zare. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Central Science*, 3(12):1337–1344, Dec. 2017.

A Coarse-grained Model

Here we use a coarse-grained model developed by DuBay group to investigate the factors influencing polymer sequences[60, 62, 40]. In the coarse-grained model simulating an irreversible step-growth copolymerization, the monomers are simplified as three connected particles: one center particle (type **1**) and two linking particles (type **2**) (Figure 7(a)).

A.1 Interaction

The non-bonded interaction between type **1** particles is represented by a modified Lennard-Jones potential (Eq (13)) with separated repulsion and attraction parts (Figure 7(a)). In all simulations, the repulsive part remains fixed, while the attractive portion varies by modifying the well depth, $\varepsilon_{\mathrm{att}}$. The ε_{AA}, ε_{AB}, and ε_{BB} refer to he attractions between **AA**, **AB**, and **BB** respectively.

$$E_{\mathrm{LJ}}(1,1') = \begin{cases} 4\epsilon_{\mathrm{att}(1,1')}[(\frac{\sigma}{r_{(1,1')}})^{12} - (\frac{\sigma}{r_{(1,1')}})^{6}] & r_0 \leq r_{(1,1')} < 2.5\sigma \\ 4\epsilon_{\mathrm{rep}(1,1')}[(\frac{\sigma}{r_{(1,1')}})^{12} - (\frac{\sigma}{r_{(1,1')}})^{6}] + C & r_{(1,1')} < r_0 \end{cases}, \tag{13}$$

The non-bonded repulsive interactions between type **2** particles is a soft short-range repulsion as described by Eq (14) and illustrated in Figure 7(a). The d_{bond} and d_{onset} refer to the bonding and interaction onset distances which are 0.2 σ and 0.3 σ respectively. The h parameter is the activation energy parameters, E_a, in the dataset, which is correlated with the activation barrier of bond forming between type **2** particles. Additional information regarding the calculation of the activation barrier and its variation with h is provided in Ref.[40].

$$E_{(2,2')} = \begin{cases} h & r < d_{\mathrm{bond}} \\ \frac{h}{2} cos(\frac{\pi(r-d_{\mathrm{bond}})}{d_{\mathrm{onset}}-d_{\mathrm{bond}}}) + \frac{h}{2} & d_{\mathrm{bond}} \leq r < d_{\mathrm{onset}} \\ 0 & r \geq d_{\mathrm{onset}} \end{cases} \tag{14}$$

All bonds are modeled as harmonic bonds with sufficiently large force constants to prevent bond breaking. The specific bond parameters are provided in our previous publication[60]. The angular potential follows the harmonic form given in Eq.(15). Here the K^{angle} denotes the potential constant. In particular, we vary the K^{angle} values for intra-monomer angles (angle α in Figure 7(a)), $K^{\mathrm{angle}}_{\mathbf{2\text{-}1\text{-}2'}}$, to modulate chain stiffness. Increasing $K^{\mathrm{angle}}_{\mathbf{2\text{-}1\text{-}2'}}$ results in a stiffer polymer backbone. The relationship between $K^{\mathrm{angle}}_{\mathbf{2-1-2'}}$ and the persistence length has been described in detail in Ref. [62]. The equilibrium bond angle, θ_0, is set to 180° for all angular potentials.

$$E_{\mathrm{angle}}(\theta_{ijk}) = K^{\mathrm{angle}}_{ijk}(\theta_{ijk} - \theta_0)^2 \tag{15}$$

A.2 Reaction

The simulations employ Langevin dynamics to all particles. The *damp* parameter in the fractional drag term of the Langevin equation is related to the solution viscosity according to Eq. (16), where m and d denote the particle mass and size, respectively:

$$\eta = \frac{m}{3\pi d \cdot damp} \tag{16}$$

Each simulation begins with a randomly distributed 1:1 mixture of **A** and **B** monomers in a cubic box with a side length of 50σ. As the reaction evolves, an irreversible bond forms when two type-**2** particles approach closer than the bonding cutoff distance d_{bond} (Figure 7(b)). The polymerization follows a step-growth mechanism, where both monomers and oligomers can react with both monomers and oligomers. Simulations are terminated when the reaction extent exceeds 0.9, meaning that over 90% of possible bonds between type-**2** particles have been formed (Figure 7(c)).

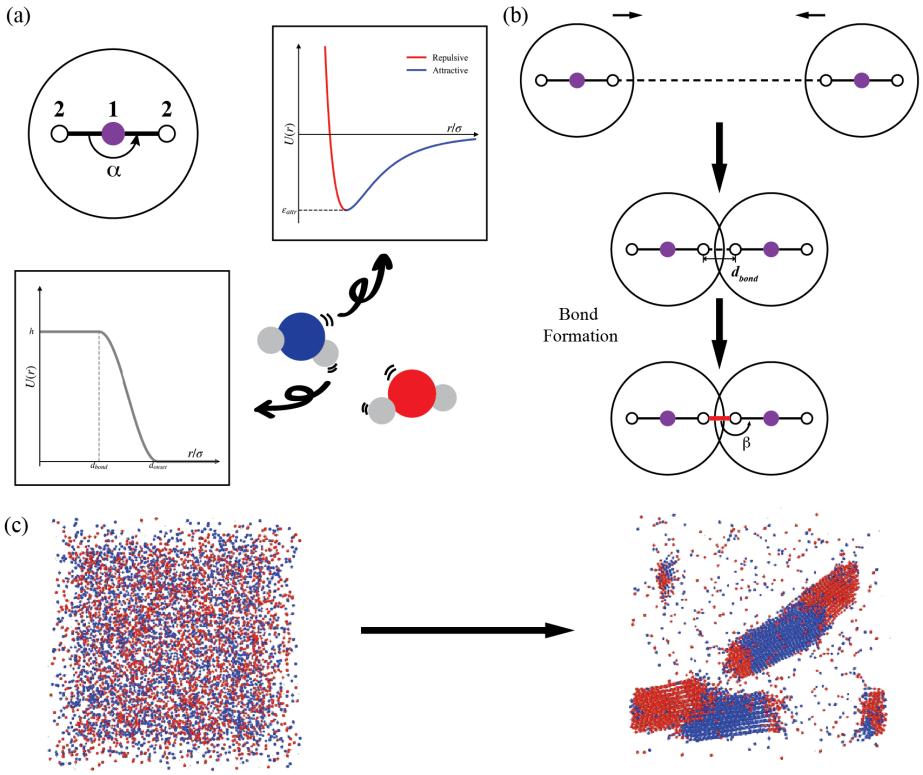


Figure 7: Illustration of the coarse-grained model for irreversible step-growth copolymerization. (a) Each monomer is represented by three particles: a central type-**1** particle, which defines the monomer type, and two type-**2** particles serving as linking sites. The angle α corresponds to the intra-monomer **2-1-2'** angle. Interactions between type-**1** particles follow a modified Lennard–Jones potential with separate repulsive and attractive parts (top right), while interactions between type-**2** particles are modeled using a short-range repulsive potential with a maximum value h (bottom left). (b)A bond forms between two type-**2** linker particles when they approach closely enough that their distance reaches d_{bond}. (c)Simulations start from a randomly distributed 1:1 mixture of **A** and **B** monomers and proceed until the reaction extent exceeds 0.9.

B Parameters in Copolymer Simulations

Table 2: Description of parameters in copolymer simulation data.

Parameter	**Description**
size	Size of the simulation box
Nmono	Number of total monomers
$damp$	Damping coefficient in Langevin Dynamics
Temp	Temperature
$\varepsilon_{\mathrm{AA}}$	Attraction between **A**-**A** monomers
$\varepsilon_{\mathrm{AB}}$	Attraction between **A**-**B** monomers
$\varepsilon_{\mathrm{BB}}$	Attraction between **B**-**B** monomers
$\varepsilon_{\mathrm{hard}}$	Repulsion between monomers
K_{A}^{angle}	Potential constant for intramolecular angle of **A** monomers
K_{B}^{angle}	Potential constant for intramolecular angle of **A** monomers
E_a^{AA}	Activation energy for the formation of **A**-**A** bonding
E_a^{AB}	Activation energy for the formation of **A**-**B** bonding
E_a^{BB}	Activation energy for the formation of **B**-**B** bonding
seq	Generated polymer sequence (an array of strings)
p_{AA}	Probability or fraction of **AA** pairs in sequence
p_{BB}	Probability or fraction of **BB** pairs in sequence
$p_{\mathrm{AA,BB}}$	Probability or fraction of **AA** or **BB** pairs in sequence
p_{AB}	Probability or fraction of **AB** pairs in sequence
block_dist	Block length distribution