Fake It Till You Make It: Multi-Physics Synthesis Breaks the Data Barrier in Chemical Language Models

Naiyu Yin
Department of Mathematics,
Lehigh University,
Bethlehem, PA 18015, USA

Ning Liu
Global Engineering and Materials Inc.,
Princeton, NJ 08540, USA

Jiuzhou Chen
Department of Mathematics,
Lehigh University,
Bethlehem, PA 18015, USA

Brian Y. Lattimer
Department of Mechanical Engineering,
Virginia Tech,
Blacksburg, VA 24061, USA

Jim Lua
Global Engineering and Materials Inc.,
Princeton, NJ 08540, USA

Yue Yu *
Department of Mathematics,
Lehigh University,
Bethlehem, PA 18015, USA

Abstract

Chemical language models (CLMs) promise transformative capabilities in polymer property prediction and design, yet their potential is hindered by the scarcity of experimental data. We present Poly4mer, a novel multi-physics synthesis based CLM framework that units neural polymer representation learning with theoretical foundations in physics. Poly4mer integrates a comprehensive group contribution method for hypothetical polymer generation with physics-based simulations that faithfully emulate experimental protocols, fabricating a rich dataset of synthetic polymer structures and structure–property relationships to establish strong priors for CLM training. This synthetic data enables training of two critical components: an encoder-decoder architecture that captures polymer semantics, and a two-phase property prediction strategy comprising supervised pretraining on synthetic data for physically consistent alignment followed by fine-tuning on experimental measurements to enhance predictive accuracy. We then architect an autoencoding system that couples predictive capability with latent decoding, enabling inverse design of polymers optimized for downstream applications through latent space exploration and structure reconstruction. By faking data with physics before reality catches up, we demonstrate that multi-physics synthesis can break the data barrier in CLMs, establishing a new paradigm for physics-grounded neural polymer discovery. Our code and pretrained models accompanying this paper are available at `https://github.com/fishmoon1234/poly4mer`.

1 Introduction

Chemical language models (CLMs) have emerged as powerful tools for molecular property prediction and generative design, leveraging SMILES-based representations and large-scale pretraining to achieve strong performance in chemistry and material science. Recent foundation models [1–4], such as SMI-TED [5], demonstrate that large encoder-decoder architectures trained on millions of

*Corresponding author. Email: `yuy214@lehigh.edu`

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

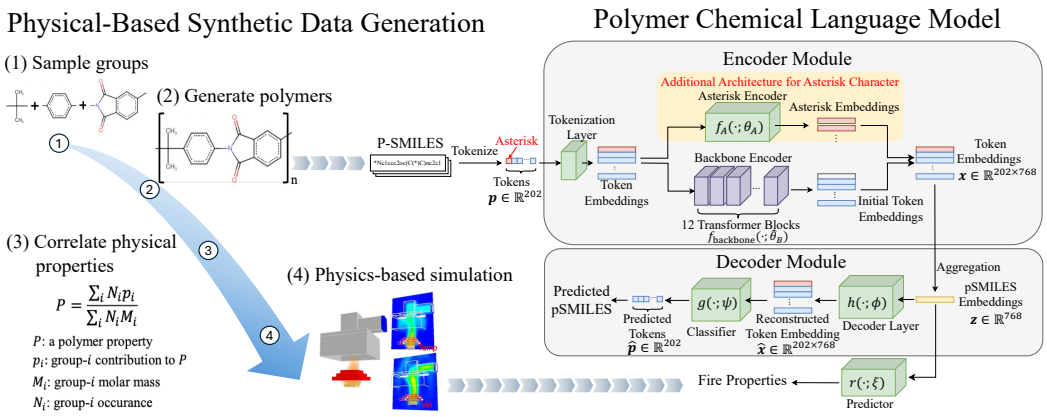

Figure 1: Illustration of the proposed physical-based data generation and CLM training.

molecules can capture chemically meaningful structure–property relationships and support diverse downstream tasks without extensive task-specific supervision.

Despite these advances, two critical bottlenecks persist. First, specialized domains such as polymer flammability prediction [6] suffer from severe data scarcity, as experimental measurements are costly and limited. Direct fine-tuning of pretrained CLMs on such small datasets often results in overfitting and poor generalization. Second, standard line notations for molecular structures pose challenges. Most CLMs rely on the "simplified molecular-input line-entry system" (SMILES) [7, 8], which is primarily designed for small molecules or monomers. Extending these representations to polymers—and enabling effective downstream fine-tuning—requires substantial retraining and large-scale polymer data, making the problem non-trivial.

To overcome these challenges, we introduce a multi-physics synthesis framework that bridges neural polymer design and first-principles physical knowledge. Our approach constructs a physics-grounded pretraining pipeline by synthesizing a large corpus of physically meaningful polymer structures and corresponding property labels using physics-based simulations. These synthetic structures provide rich semantic information about polymers, while the simulation-derived labels serve as strong physical priors for CLMs. After pretraining, models are fine-tuned on limited experimental data for optimal predictive accuracy. Furthermore, we architect a unified autoencoding system that jointly trains the CLM encoder module, decoder module, and property predictor, enabling inverse polymer design through principled latent space exploration and structure reconstruction. We demonstrate that physics-informed synthesis effectively breaks the data barrier in CLMs. Our contributions include:

- We establish a systematic group contribution method for generating hypothetical polymers that bridges data-driven modeling and first-principles physical knowledge.
- We introduce a principled multi-physics synthesis framework to data-scarce learning by embedding domain knowledge directly into language model training, enabling robust generalization beyond pure data-driven models.
- We design a new strategy for introducing new tokens into pretrained language models via encoder-decoder architecture expansion, leveraging monomer representations to capture complex polymer semantics.
- We architect an autoencoding system coupling predictive modeling with generative design, enabling inverse polymer design via latent-space exploration and structure reconstruction for targeted applications.

2 Related Work

Chemical Language Models. CLMs leverage transformer architectures for molecular property prediction by encoding small-molecule structures as SMILES sequences [1, 4, 5, 9–15]. While these models achieve strong performance through large-scale pretraining, their effectiveness is fundamentally constrained by the availability of experimental data. Recent work has extended CLMs to polymers [14, 15] by introducing additional tokens in SMILES to represent covalent bonds and

repeating units. However, the increased complexity of polymer structures exacerbates the data scarcity problem: unlike small-molecule databases [16] with millions of labeled compounds, existing polymer datasets typically contain only tens to hundreds of experimentally characterized samples [17].

Generative Polymer Design. Traditional polymer discovery relies on costly experimental trial-and-error. To accelerate this process, recent methods have explored variational autoencoders for molecular generation [18, 19], graph neural networks for structure synthesis [20, 21], and physics-informed neural networks for incorporating physical constraints [22]. However, none of these approaches systematically address the polymer data bottleneck through physics-based synthesis.

Physics-Guided Machine Learning. Physics-guided machine learning [23–31] integrates domain knowledge into data-driven models to mitigate data scarcity challenges. Examples in polymer modeling include RNNs for predicting viscoelastic behavior [32], equivariant graph networks for capturing molecular interactions [33], and handcrafted physical fingerprints for polymer characterization [34]. Despite these advances, a systematic framework that couples physics-based synthetic data generation with CLM training remains largely unexplored.

3 Physics-Based Chemical Language Model Training

We present an overview of the proposed Poly4mer framework, as illustrated in Figure 1.

Physics-Based Synthetic Data Generation. To obtain structurally valid and physically meaningful polymers, we adopt the group contribution (GC) method [35–38]. This method estimates polymer properties based on the frequency of functional groups and their contributions. We identify 48 functional groups capable of generating 180 polymers with available micro-combustion calorimeter (MCC) data (Step 1, Figure 1). A global optimization routine determines the contribution of each group, and the resulting MCC property data are used as inputs to the Fire Dynamics Simulator (FDS) [39], a physics-based reduced-order model for predicting fire properties. Although the GC method simplifies molecular interactions and introduces some accuracy limitations, it is computationally efficient and enforces physical constraints through admissible structures and simulation-based properties. We leverage it to construct a large synthetic polymer dataset as follows: (1) sample base functional groups and combine them into homopolymers with two open valences; (2) assemble these homopolymers into copolymers and terpolymers, and canonicalize the resulting structures to obtain unambiguous polymer SMILES; (3) use the GC method and FDS simulations (Steps 3–4, Figure 1) to estimate four flammability metrics: *time to ignition (T_{ig})*, *peak heat release rate (pHRR)*, *smoke extinction area (SEA)*, and *carbon monoxide yield (CO)*. This pipeline yields $\sim$120k admissible synthetic polymer structures with physically consistent property labels, which serve as priors for learning complex structure–property relationships in our Poly4mer predictor.

Polymer Representation Learning. Polymer SMILES introduces an additional symbol '*' to represent open bonds for polymer connectivity [14]. This creates challenges for CLMs pretrained on monomers and lack both this token and the associated polymer semantics. Therefore, state-of-the-art CLMs, such as MolFormer [4] and SMI-TED [5], still cannot encode or generate '*' in SMILES nor handle polymer structures. The only exception is TransPolymer [15], which incorporates '*' for property prediction. However, it does not possess the capability for polymer generation. To provide a polymer CLM coupling predictive modeling with generative design, we propose an *encoder–decoder foundation model tailored for polymers* (Figure 1), extending the SMI-TED architecture [5]. Given a polymer SMILES tokenized as $\boldsymbol{p} = [p_1, p_2, \ldots, p_L] \in \mathbb{R}^{L=202}$, the encoder maps it to a token embedding $\boldsymbol{x} \in \mathbb{R}^{202\times 768}$ and a polymer-level embedding $\boldsymbol{z} \in \mathbb{R}^{768}$. The latent representation $\boldsymbol{z}$ captures essential polymer semantics and is used for token reconstruction and property prediction.

Then, we separately train the encoder and decoder modules on the synthetic polymer dataset. The encoder consists of two components: (i) a backbone encoder $f_{\text{backbone}}(\cdot;\hat{\theta}_B)$ pretrained on monomers, and (ii) an asterisk-specific encoder $f_A(\cdot;\theta_A)$ to handle the '*' token. Given $\boldsymbol{p}$, the backbone produces initial embeddings $\boldsymbol{x} = f_{\text{backbone}}(\boldsymbol{p};\hat{\theta}_B)$, while f_A generates embeddings for positions corresponding to the asterisk token, replacing the respective rows in $\boldsymbol{x}$. The full encoder is denoted as $f = (f_{\text{backbone}}, f_A)$, yielding $\boldsymbol{x} = f(\boldsymbol{p};\hat{\theta}_B,\theta_A)$. To train the encoder, we attach a surrogate classifier $g_s(\cdot;\psi_s)$ that predicts the token indices from $\boldsymbol{x}$, and solve:

$$\hat{\theta}_A, \hat{\psi}_s = \arg\min_{\theta_A,\psi_s} \mathcal{L}_{\text{pred}}\Big(\boldsymbol{p}, g_s\big(f(\boldsymbol{p};\theta_A,\hat{\theta}_B);\psi_s\big)\Big), \tag{1}$$

where $\mathcal{L}_{\text{pred}}(\cdot,\cdot)$ denotes the cross-entropy loss. The polymer embedding is then obtained by aggregating over the L token dimensions, i.e., $\boldsymbol{z} = \sum_{l=1}^{L} \boldsymbol{x}[l,:]$.

The *decoder module* consists of a decoder layer $h(\cdot;\phi)$ and a classifier $g(\cdot;\psi)$, with ϕ and ψ denoting their respective parameters. h reconstructs the token embedding from the polymer embedding, while g predicts token indices from the reconstructed embedding. We train both components jointly by minimizing a weighted sum of the reconstruction loss $\mathcal{L}_{\text{rec}}$ and the prediction loss $\mathcal{L}_{\text{pred}}$:

$$\hat{\psi}, \hat{\phi} = \arg\min_{\psi,\phi} \lambda \mathcal{L}_{\text{rec}}(\boldsymbol{x}, \boldsymbol{z}; \phi) + \mathcal{L}_{\text{pred}}\Big(\boldsymbol{p}, g\big(h(\boldsymbol{z};\phi);\psi\big)\Big), \tag{2}$$

whereby $\mathcal{L}_{\text{rec}}(\boldsymbol{x}, \boldsymbol{z}; \phi) \coloneqq \|\boldsymbol{x} - h(\boldsymbol{z};\phi)\|_F^2$, λ is the coefficient for the reconstruction loss. It is worth noting that including the decoder module is crucial for extending Poly4mer beyond property prediction to inverse polymer design, a capability that state-of-the-art single-purpose CLMs such as TrinityLLM [1] lack.

Two-Phase Property Predictor Training. Leveraging the physics-guided training of encoder-decoder architecture, we obtain a contextualized polymer embedding $\boldsymbol{z}$ for each polymer, which serves as the input to a property predictor for flammability metrics. For each target property y, we train a shallow MLP $r(\cdot;\xi)$ by minimizing the prediction loss:

$$\hat{\xi} = \arg\min_{\xi} \mathcal{L}_{\text{pred}}\big(y, r(\boldsymbol{z};\xi)\big). \tag{3}$$

To leverage the physics-consistent synthetic dataset generated via the GC method, we adopt a two-phase training strategy. Phase I pre-trains the predictors on a large synthetic dataset of approximately 120k polymers generated using the GC method. This stage enables the predictors to inherit physical priors from the simulation-based data, providing a well-informed initialization. Phase II fine-tunes the predictors on an experimental dataset comprising 55 polymers. Since these experimental measurements of flammability metrics represent high-fidelity data, Phase II is expected to further enhance predictive accuracy.

4 Polymer Discovery

With the generative modeling and fire property prediction capabilities, the trained Poly4mer model enables polymer discovery as a downstream application. Given a trained CLM (f, h, g, r) with parameters $(\hat{\theta}_A, \hat{\theta}_B, \hat{\phi}, \hat{\psi}, \hat{\xi}_{tig}, \hat{\xi}_{pHRR}, \hat{\xi}_{SEA}, \hat{\xi}_{CO})$, our objective is to discover polymers with long time to ignition, low peak heat release rate, low smoke extinction area, and low carbon monoxide yield. To this end, we define the property optimization objective:

$$\mathcal{L}_{\text{obj}} \coloneqq -\gamma_{tig} r(\boldsymbol{z};\hat{\xi}_{tig}) + \gamma_{pHRR} r(\boldsymbol{z};\hat{\xi}_{pHRR}) + \gamma_{SEA} r(\boldsymbol{z};\hat{\xi}_{SEA}) + \gamma_{CO} r(\boldsymbol{z};\hat{\xi}_{CO}). \tag{4}$$

Here, γ_* denotes the corresponding penalty coefficient, so users can assign weights according to their preferences on the desired properties. During the discovery phase, we optimize the polymer embedding by minimizing $\mathcal{L}_{\text{obj}}$, and subsequently reconstruct the corresponding polymer tokens using the decoder:

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z}} \mathcal{L}_{\text{obj}}, \quad \boldsymbol{p}^* = g\big(h(\boldsymbol{z}^*;\hat{\phi});\hat{\psi}\big). \tag{5}$$

5 Experiment

To evaluate the effectiveness of the trained CLM for polymer discovery, we first assess its performance in polymer reconstruction and property prediction, followed by case studies demonstrating the feasibility of the proposed framework.

Polymer Reconstruction. We measure reconstruction accuracy on 125,743 polymer SMILES, including 55 experimental cases and 125,688 synthetically generated ones. The trained CLM achieves 99.89% accuracy on synthetic polymers and 98.18% on experimental polymers, indicating high fidelity in reconstructing polymer structures from the encoder–decoder architecture.

Property Prediction. We first train the predictors on physics-based synthetic data and obtain decent relative Mean Squared Error (MSE) values of 6.7%, 12.4%, 5.7%, and 8.0% for t_{ig}, pHRR, SEA, and CO. We then finetune and evaluate the property prediction performance on experimental data, with test dataset results summarized in Table 1. Empirically, our method outperforms state-of-the-art baselines across all four fire-related properties, highlighting the effectiveness of the proposed predictor and the benefits of the two-phase training strategy.

Generative Polymer Design. We illustrate the generative polymer design process using polyvinyl chloride (PVC) as an exemplar starting structure (Figure 2). The optimization penalty parameters

Table 1: Fire properties prediction performance on experimental data in terms of relative MSE (%) against state-of-the-art baselines.

Methods	Fire Properties			
	t_{ig} ↓	pHRR ↓	SEA ↓	CO ↓
MoLFormer [4]	34.04 ± 22.07	62.01 ± 19.94	41.11 ± 14.70	-
TransPolymer [15]	31.78 ± 5.81	77.88 ± 76.64	38.91 ± 16.95	-
polyBERT [14]	41.39 ± 5.09	61.42 ± 11.26	51.28 ± 10.00	-
TrinityLLM [1]	16.65 ± 5.72	33.64 ± 2.06	22.35 ± 12.24	18.58 ± 8.28
Poly4mer (ours)	**16.32** ± 6.20	**21.33** ± 3.23	**17.47** ± **4.19**	**16.13** ± **2.97**

are taken as: $\gamma_{tig} = 0.1$, $\gamma_{pHRR} = \gamma_{SEA} = \gamma_{CO} = 1$. Through iterative optimization of the latent embedding and subsequent decoding at each step, the algorithm progressively refines the polymer representation and outputs all discovered candidate polymer SMILES (see the middle plot of Figure 2). After passing the candidate polymer SMILES through a canonicalizer, feasible polymer structures transition from the initial $*CC(*)Cl$ to $*CCC*$, then to $*CCC(*)Cl$, and ultimately reaching a better structure $*CCC(*)C$, as shown in the right plot of Figure 2.

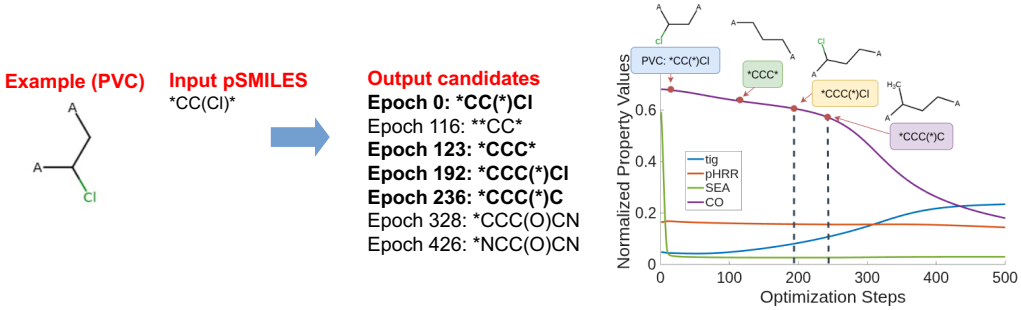

Figure 2: An illustration of our polymer design procedure. We begin with the polymer PVC, represented by the SMILES $*CC(*)Cl$. Its latent embedding is optimized by minimizing the training objective in Eq. (4), and the resulting optimal embeddings are decoded to generate new polymers. The line plot shows the optimization trajectory of PVC.

6 Conclusion and Discussion

We present Poly4mer, a multi-physics synthesis framework that addresses data scarcity in chemical language models through physics-based synthetic data generation. By integrating group contribution theory with physics-based simulations, Poly4mer establishes meaningful polymer semantics and strong physical priors for CLM training. By using the physics-based synthetic data to pretrain encoder/decoder modules together with a fire property predictor, our Poly4mer achieves state-of-the-art performance in fire property prediction while also enabling inverse polymer design. This success illustrates a "fake it till you make it" paradigm, where physics-grounded synthesis effectively overcomes data limitations in specialized domains and opens new pathways for accelerated polymer discovery.

We also point out that Poly4mer has several limitations. First, we generate physics-based synthetic polymer SMILES in canonicalized form and train Poly4mer only on these canonical SMILES. This restricts its generalization to other encoding forms (for example, greedy encoding) and prevents the model from distinguishing chemically equivalent polymers, as they share the same canonical SMILES. Incorporating polymer structure from alternative modalities could be a promising future direction. Second, we can only qualitatively claim that Poly4mer captures meaningful polymer semantics, based on empirical results showing that latent representations are effective for both polymer generation and property prediction. Mechanistic interpretability studies are needed to verify whether the model truly captures polymer chemistry. Finally, the current version focuses on polymer fire properties as an exemplar demonstration. Extending Poly4mer to a broader range of polymer properties, such as the mechanical and fracture properties, would be a key goal for future work.

Acknowledgments and Disclosure of Funding

This work is supported by the Office of Naval Research (ONR) under grant numbers N68335-24-C-0123 and N68335-25-C-0210. Portions of this research were conducted on Lehigh University's Research Computing infrastructure partially supported by NSF Award 2019035.

References

[1] Ning Liu, Siavash Jafarzadeh, Brian Y Lattimer, Shuna Ni, Jim Lua, and Yue Yu. Harnessing large language models for data-scarce learning of polymer properties. *Nature Computational Science*, 5(3):245–254, 2025.

[2] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

[3] Chang Liao, Yemin Yu, Yu Mei, and Ying Wei. From words to molecules: A survey of large language models in chemistry. *arXiv preprint arXiv:2402.01439*, 2024.

[4] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[5] Emilio Vital Brazil, Eduardo Soares, Victor Yukio Shirasuna, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. Smi-ted: A large-scale foundation model for materials and chemistry. In *AI for Accelerated Materials Design-ICLR 2025*.

[6] Richard E Lyon, Marc L Janssens, et al. Polymer flammability. Technical report, United States. Federal Aviation Administration. Office of Aviation Research, 2005.

[7] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[8] David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101, 1989.

[9] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[10] Dongyu Xue, Han Zhang, Dongling Xiao, Yukang Gong, Guohui Chuai, Yu Sun, Hao Tian, Hua Wu, Yukun Li, and Qi Liu. X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv*, pages 2020–12, 2020.

[11] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436, 2019.

[12] Hyunseob Kim, Jeongcheol Lee, Sunil Ahn, and Jongsuk Ruth Lee. A merged molecular representation learning for molecular properties prediction with a web-based service. *Scientific Reports*, 11(1):11028, 2021.

[13] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

[14] Christopher Kuenneth and Rampi Ramprasad. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature communications*, 14(1):4099, 2023.

[15] Changwen Xu, Yuyang Wang, and Amir Barati Farimani. Transpolymer: a transformer-based language model for polymer property predictions. *npj Computational Materials*, 9(1):64, 2023.

[16] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[17] Richard E Lyon, Michael T Takemori, Natallia Safronava, Stanislav I Stoliarov, and Richard N Walters. A molecular basis for polymer flammability. *Polymer*, 50(12):2608–2617, 2009.

[18] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31, 2018.

[19] Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International conference on artificial neural networks*, pages 412–422. Springer, 2018.

[20] Gang Liu, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph diffusion transformers for multi-conditional molecular generation. *Advances in Neural Information Processing Systems*, 37:8065–8092, 2024.

[21] Rocío Mercado, Tobias Rastemo, Edvard Lindelöf, Günter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum. Graph networks for molecular design. *Machine Learning: Science and Technology*, 2(2):025023, 2021.

[22] Sara Ibrahim Omar, Chen Keasar, Ariel J Ben-Sasson, and Eldad Haber. Protein design using physics informed neural networks. *Biomolecules*, 13(3):457, 2023.

[23] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[24] Ivan Malashin, Vadim Tynchenko, Andrei Gantimurov, Vladimir Nelyub, and Aleksei Borodulin. Physics-informed neural networks in polymers: A review. *Polymers*, 17(8):1108, 2025.

[25] Eric Inae, Yuhan Liu, Yihan Zhu, Jiaxin Xu, Gang Liu, Renzheng Zhang, Tengfei Luo, and Meng Jiang. *Modeling Polymers with Neural Networks*. American Chemical Society, 2025.

[26] Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan S Read, Jacob A Zwart, Michael Steinbach, and Vipin Kumar. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions on Data Science*, 2(3):1–26, 2021.

[27] Lanyi Wang, Shun-Peng Zhu, Changqi Luo, Ding Liao, and Qingyuan Wang. Physics-guided machine learning frameworks for fatigue life prediction of am materials. *International Journal of Fatigue*, 172:107658, 2023.

[28] Yuyao Chen and Luca Dal Negro. Physics-informed neural networks for imaging and parameter retrieval of photonic nanostructures from near-field data. *APL Photonics*, 7(1), 2022.

[29] Ning Liu, Xuxiao Li, Manoj R Rajanna, Edward W Reutzel, Brady Sawyer, Prahalada Rao, Jim Lua, Nam Phan, and Yue Yu. Deep neural operator enabled digital twin modeling for additive manufacturing. *arXiv preprint arXiv:2405.09572*, 2024.

[30] Siavash Jafarzadeh, Stewart Silling, Ning Liu, Zhongqiang Zhang, and Yue Yu. Peridynamic neural operators: A data-driven nonlocal constitutive model for complex material responses. *Computer Methods in Applied Mechanics and Engineering*, 425:116914, 2024.

[31] Ning Liu, Siavash Jafarzadeh, and Yue Yu. Domain agnostic fourier neural operators. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Bao Qin and Zheng Zhong. A physics-guided machine learning model for predicting viscoelasticity of solids at large deformation. *Polymers*, 16(22):3222, 2024.

[33] Rui Feng, Huan Tran, Aubrey Toland, Binghong Chen, Qi Zhu, Rampi Ramprasad, and Chao Zhang. Polyget: Accelerating polymer simulations by accurate and generalizable forcefield with equivariant transformer. *arXiv preprint arXiv:2309.00585*, 2023.

[34] Huan Doan Tran, Chiho Kim, Lihua Chen, Anand Chandrasekaran, Rohit Batra, Shruti Venkatram, Deepak Kamal, Jordan P Lightstone, Rishi Gurnani, Pranav Shetty, et al. Machine-learning predictions of polymer properties with polymer genome. *Journal of Applied Physics*, 128(17), 2020.

[35] Leonidas Constantinou and Rafiqul Gani. New group contribution method for estimating properties of pure compounds. *AIChE Journal*, 40(10):1697–1710, 1994.

[36] Al L Lydersen. Estimation of critical properties of organic compounds. *Univ. Wisconsin Coll. Eng., Eng. Exp. Stn. Rep. 3*, 1955.

[37] KM Klincewicz and RC Reid. Estimation of critical properties with group contribution methods. *AIChE Journal*, 30(1):137–142, 1984.

[38] Kevin G Joback and Robert C Reid. Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications*, 57(1-6):233–243, 1987.

[39] Tzu-Sheng Shen, Yu-Hsiang Huang, and Shen-Wen Chien. Using fire dynamic simulation (fds) to reconstruct an arson fire scene. *Building and environment*, 43(6):1036–1045, 2008.