**De-Novo Generation of Synthetic Copolymers with Graph-to-String Variational Autoencoder**

Gabriel Vogela, Paolo Sortinob, Jana Marie Webera,\*

aDelft Bioinformatics Lab, **Departmentof Intelligent Systems, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands**

bDeptartment of Computer Engineering, University of Palermo, Palermo, Italy

*j.m.weber@tudelft.nl*

Abstract

Synthetic polymers are a key material class in many applications with importance to our society, such as plastics and novel energy materials. This causes an increasing demand for new synthetic polymers with superior properties. However, due to their complex hierarchical and highly versatile structure, the theoretical design space for polymers is large. Generative polymer design is a promising approach to accelerate polymer materials discovery and reduce experimental screening costs. Yet, generative polymer design is still in its infancy with most previous approaches focusing on generating the repeating units of polymers without additional structural information. Here, we present our recently published graph-to-string variational autoencoder (VAE) that builds upon a recent polymer graph representation including stoichiometries and chain architectures of monomer ensembles. The model enables de-novo generation of copolymer structures including the monomer stoichiometry and chain architecture. In this work, we demonstrate three different sampling strategies to generate novel polymers, namely sampling from noise, around a seed polymer, and through interpolation between two polymers. All methods generate novel polymers not found in the training data, mainly through new monomer combinations and novel monomer chemistries. Sampling around seed molecules and interpolation between molecules illustrate how variations in the latent space change the generated polymers.

**Keywords**: generative molecular design, synthetic polymers, variational autoencoder, transformers, graph neural network

* 1. Introduction

De-novo generative design of molecules is an increasingly popular approach to intelligently design molecules in-silico, reducing experimental screening time and costs. Generative models are commonly trained on a large corpus of molecular data and facilitate the generation of promising novel molecular structures which can then be synthesized in the lab. Yet, compared to the small molecule world, generative synthetic polymer design is still in its infancy. One reason is the overall small amount of available, accessible, and sufficiently detailed polymer data (Amamoto, 2022; C. Kim et al., 2018; Otsuka et al., 2011). Second, unlike for small molecules, the definition of a machine-readable data representation that accurately corresponds to the final polymer material is difficult (Amamoto, 2022; Hatakeyama-Sato, 2022; Yan & Li, 2023). This results from the stochastic nature of polymers and their different structural levels reaching from monomer chemistry over monomer composition and stoichiometry to chain architecture and linking structure, visualized in Figure 1.

Figure 1: Structural levels of polymers: Monomers, combination of monomers and stoichiometry, chain architecture, and chain linking structure

Most previous works on generative design of polymers focuses mainly on generating single polymer repeating units with little or no information about the higher-order structural levels (S. Kim et al., 2023; Ma & Luo, 2020). In this work, we demonstrate our recently published graph-to-string variational autoencoder (VAE) (Vogel et al., 2023), a generative model that has been trained on polymer representations that include the stoichiometry of the monomer ensemble and the chain architecture of the polymer (Aldeghi & Coley, 2022). The model encodes the training data to a continuous numerical latent space that can be used to generate novel copolymers including the monomer stoichiometry and chain architecture. We demonstrate three different sampling strategies in the model’s latent space, namely sampling from noise, around a seed molecule and through interpolation between known molecules. Finally, we discuss the implications of the sampling results for future work.

* 1. Methods
		1. Representations and Data

The model is trained on the dataset published by (Aldeghi & Coley, 2022), which is built upon the polymer space of (Bai et al., 2019). The dataset comprises combinations of eight A-monomers with 682 B-monomers with the stoichiometries 1:1, 1:3, and 3:1 and three chain architectures (alternating, random, block), leading to a dataset of 42966 copolymers. Additionally, we augment the dataset by allowing B-B copolymers with the same selection of stoichiometries and chain architectures, as described in Vogel et al. 2023. We work with the polymer graph representation introduced by Aldeghi & Coley, 2022 together with an equivalent string format, both including the information about stoichiometry and chain architecture. Further information regarding the polymer representations can be found in (Vogel et al., 2023).

* + 1. Model

We use our recently published model to encode the polymer graph to a latent representation $z\_{x}$ with latent dimension 32 and to decode it to the corresponding polymer string. As shown in Figure 2, in the encoding step the model utilizes a graph neural network to learn the mean $μ\_{x}$ and standard deviation $σ\_{x}$ of the data which is then reparametrized using a gaussian prior. The decoder is based on the transformer architecture, using the latent representation through encoder-decoder attention and through concatenation with the token embeddings. More details on the model architecture and training hyperparameters are provided in the original paper (Vogel et al., 2023).



Figure 2: Simplified overview of Graph-to-string VAE model for synthetic polymer design (Vogel et al., 2023). Polymer graphs are encoded in a wDMPNN (Aldeghi & Coley, 2022) to learn $μ\_{x}$and $σ\_{x}$ which are then reparametrized to the latent representation$z\_{x}$. The latent representation is fed to a Transformer decoder that decodes the equivalent polymer strings.

* + - 1. Sampling: from noise, around seed, and using interpolation

In this work, we demonstrate de-novo generative design of synthetic polymers using three different sampling techniques. First, novel polymers can be generated by sampling latent vectors$ $from Gaussian noise $z\_{n\\_gaussian}$and decoding these using the trained model. Second, we sample novel polymers around a given seed molecule, i.e. by repeatedly adding noise$n$(of latent dimension 32) elementwise to a seed latent vector$z\_{s}$**,** leading to $z\_{s+n}$.Lastly, we interpolate between molecules, i.e. between two latent codes for two known molecules $z\_{m1}$ and $z\_{m2}$. One can take the mean or interpolate in a stepwise manner. In this work, we create ten interpolated latent codes $z\_{i, m1\rightarrow m2}$, by interpolating elementwise in equidistant steps between the two latent codes of the known molecules. For instance, let latent dimension one be 0.2 for $z\_{m1}$ and 0.75 for $z\_{m2}$**,** then the interpolated values of latent dimension one $z\_{1,i, m1\rightarrow m2}$, would be $\{0.25, 0.3, 0.35, …, 0.7\}$ for the respective interpolated latent vectors $z\_{i, m1\rightarrow m2}, i\in \{1, 2, 3, …, 10\}$.

* 1. Results and Discussion

In the following, we show the results of de-novo generation of synthetic polymers focusing on the three proposed sampling techniques. Moreover, we discuss how the monomer chemistries, stoichiometry, and chain architecture change when changing the latent codes of polymers.

* + 1. Sampling from gaussian noise

We use the model to sample 16000 polymers from gaussian noise leading to polymers with 77 % novelty (percentage of polymers not in training set), 40 % diversity (percentage of unique polymers), and >99 % validity (polymer string corresponds to a valid molecule). Further, when investigating the novelty per structural category, we observe 10.9 % novel B monomer chemistries and 2.3 % novel A monomer chemistries. We observe no novel stoichiometries and chain architectures, meaning that the model is currently limited to generating the classes found in the training data. Thus, the overall novelty is introduced through novel monomers and novel combinations of structural components. This indicates that structural categories with a higher variation in the training data lead to higher novelty during sampling, demonstrating the necessity to diversify the dataset in all structural categories in future efforts.

* + 1. Sampling around a seed molecule

Figure 3 shows example molecules when sampling 512 times around a seed molecule by repeatedly adding noise to the seed latent code $z\_{s}$**.** We observe variation from the seed molecule in all structural levels, i.e. monomer A, monomer B, stoichiometry, and chain architecture. Also, there are several instances of sampling the seed molecule itself (in this example 3 out of 512). The level of noise can be varied, with less noise leading to less changes from the seed molecule. Figure 3b reveals that, for this example, the structural level that is varied the most when adding noise is monomer B, followed by monomer A, stoichiometry, and the chain architecture. The higher variation in monomer B is expected, since the diversity in the training data is the highest. As already mentioned in Section 3.1, we also find novel B-monomers and few novel A-monomers (not present in the dataset) indicated as the red hatched bar in Figure 3b.



Figure 3: Sampling around a seed molecule. (a) Different structural categories (monomer A, monomer B. stoichiometry, and chain architecture) are varied. (b) The bar chart shows how often out of 512 samples the respective structural categories are varied from the seed molecule. The red (hatched) bars indicate novelty in the respective structural category.

* + 1. Interpolation between molecules

Figures 4a and 4b show examples of the interpolation path between two copolymers, where monomer A is depicted in the upper part and monomer B in the lower part. The interpolation pathway reveals that the different structural levels transform stepwise from the start molecule to the end molecule. Furthermore, we observe that the stoichiometry and chain architecture change less frequently, yet altering the latent code sufficiently leads to a structural change in the decoded polymer. On the contrary, the monomers change already for smaller variations of the latent vector throughout the interpolation, especially monomer B, as expected based on the diversity of monomer B in the dataset. Notably, all polymer intermediates in Figure 4 are not found in our dataset and are novel, valid copolymers. The novelty of intermediates in Figure 4a is only due to novel combinations, meaning that all monomers in this interpolation path are found in the dataset. In the example in Figure 4b we also observe novel monomers, in line with novel A and B monomers found when sampling around seed molecules (see Figure 3b).

In both examples, we see that heteroatoms (no C or H) and functional groups play a key role how the monomers change. Heteroatoms like S, F, Cl, N, and O determine whether polymers are encoded in similar regions in the latent space. For instance, in Figure 4b, the start molecule contains four oxygen atoms, while the end molecule does not contain any. Throughout the interpolation path, the oxygen atoms are disappearing in the decoded molecules. Similarly, the end molecule contains a chlorine atom, while the start molecule does not. In the middle of the interpolation part a chlorine atom appears, changing its position in the polymer between the two monomers until reaching the end molecule. This behavior is desirable, since functional groups such as chloride or nitro groups determine the properties of the polymer and therefore polymers with the same functional groups should be close in latent space.

Furthermore, all sampled polymers are conjugated copolymers, in line with the chemical space that the model is trained on which consists of conjugated copolymers used as photocatalysts (Bai et al., 2019).



Figure 4: Two examples of interpolation paths between two polymers showing sequential changes in monomer A (upper molecule), monomer B, stoichiometry and chain architecture. Changing parts of the polymer compared to the previous step are highlighted in blue circles. The green number indicates for how many steps this polymer is decoded on the interpolated path.

* 1. Conclusions

We developed a model for de-novo generative design of synthetic copolymer structures including monomer chemistries, stoichiometry and chain architecture. Our approach increases the level of information about the polymer, enabling the generation of polymer structures that represent the polymer material more accurately. Here, we demonstrate with three sampling techniques how the model can be used to generate novel copolymers and show that novelty is introduced by new combinations of known structural levels, novel monomer chemistries and the combination of the two. Moreover, in this work we show that polymers with similar functional groups and heteroatoms are encoded in similar regions of the latent space. Prospectively, with an increased diversity of the dataset on all structural levels, we envision a model that can generate novel and diverse polymer structures, including, for instance, novel stoichiometries.

References

Aldeghi, M., & Coley, C. W. (2022). A graph representation of molecular ensembles for polymer property prediction. Chemical Science, 13(35), 10486–10498.

Amamoto, Y. (2022). Data-driven approaches for structure-property relationships in polymer science for prediction and understanding. Polymer Journal, 54(8), 957–967.

Bai, Y., Wilbraham, L., Slater, B. J., Zwijnenburg, M. A., Sprick, R. S., & Cooper, A. I. (2019). Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory. Journal of the American Chemical Society, 141(22), 9063–9071.

Hatakeyama-Sato, K. (2022). Recent advances and challenges in experiment-oriented polymer informatics. Polymer Journal, 1–15.

Kim, C., Chandrasekaran, A., Huan, T. D., Das, D., & Ramprasad, R. (2018). Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. The Journal of Physical Chemistry C, 122(31), 17575–17585.

Kim, S., Schroeder, C. M., & Jackson, N. E. (2023). Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. ACS Polymers Au.

Ma, R., & Luo, T. (2020). PI1M: A Benchmark Database for Polymer Informatics. Journal of Chemical Information and Modeling, 60(10), 4684–4690.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011). PoLyInfo: Polymer Database for Polymeric Materials Design. 2011 International Conference on Emerging Intelligent Data and Web Technologies, 22–29.

Vogel, G., Sortino, P., & Weber, J. (2023, November 3). Graph-to-String Variational Autoencoder for Synthetic Polymer Design. AI for Accelerated Materials Design - NeurIPS 2023 Workshop.

Yan, C., & Li, G. (2023). The Rise of Machine Learning in Polymer Discovery. Advanced Intelligent Systems, n/a(n/a), 2200243.