Supplementary Materials

Ensuring Physicochemical Fidelity of Generated Polymers with PoGE

Ivan Bespalov[a,b], Ivan Zlobin[a,c], Mark Schneider[b], Gleb Averochkin[c], Eugeny Alexandrov[c]

[a]Kurnakov Institute of General and Inorganic Chemistry, Russian Academy of Sciences, Leninskii Prosp. 31, 119991 Moscow, Russian Federation
[b]Lomonosov Moscow State University, Leninskie Gory, 1(3), 119991 Moscow, Russian Federation
[c]Center NTI "Digital Materials Science: New Materials and Substances", Bauman Moscow State Technical University, 2S1 Baumanskaya St., 5/1, 105005 Moscow, Russian Federation

1. Key Cheminformatics Concepts

In cheminformatics, molecular structures are represented and analyzed using specialized representations and heuristics designed to capture chemical relevance. We briefly define the key concepts used in this work:

- Extended Connectivity Fingerprints[1] (ECFP): ECFP is a widely adopted method for generating fixed-length binary vectors that encode the local chemical environment around each atom in a molecule. Starting from atomic properties (e.g., element type, degree, charge), ECFP iteratively expands the neighborhood of each atom up to a specified radius (typically 2–3 bonds), hashing combinations of atom environments into bit vectors. These fingerprints serve as structural descriptors that capture topological similarity and are commonly used as input features in ML models for property prediction, virtual screening, and molecular generation.
- RDKit's SMILES-to-MOL Conversion for Validity Assessments:
- SMILES (Simplified Molecular Input Line Entry System) is a string-based notation for representing molecular structures. RDKit is a popular open-source cheminformatics toolkit that parses SMILES strings into structured 3D-aware molecular graphs (MOL objects). In generative modeling, generated SMILES strings are often invalid due to syntax errors or unphysical bonding (e.g., pentavalent carbon). By attempting conversion to MOL format, we assess chemical validity: only molecules that successfully parse into a chemically plausible structure are considered valid outputs. The ratio of valid molecules is a standard metric for evaluating the quality of molecular generative models.

- BRICS[2] Fragments: BRICS (Breakable Rules for Intelligent Chemical Synthesis) is a rule-based fragmentation algorithm that decomposes molecules at pre-defined, synthetically plausible bond types (e.g., amide, ester, ether linkages). Each fragment retains chemically meaningful substructures that correspond to common small-molecule synthetic building blocks. BRICS fragments are used here to analyze modularity and reusability of generated molecules, enabling comparisons with known retrosynthetic pathways and facilitating interpretable analysis of scaffold diversity.
- Bemis–Murcko[3] Scaffolds: A Bemis–Murcko scaffold is the core ring system and connecting linkers of a molecule, with all side chains removed. It captures the essential topological framework of a compound and is used to classify molecules by structural class. It is usually analyzed as a metric for avoiding repetitive or overly similar outputs in generative models, and for assessing coverage of chemically relevant chemical space for the domain of small molecules.

2. Training Data Acquisition

2.1 PolyTAO corpus generation

PolyTAO is designed as a conditional generative model for polymer structures, requiring specific descriptor values (Table S1) as input for structure generation. To enable unconditional generation - that is, the generation of polymer structures without predefined constraints - the model must first be adapted to accommodate this objective.

Table S1. PolyTAO input descriptors

Descriptor Name	Descriptor detailed name	Descriptor type
MolWt	Molecular weight of monomer	real number
HeavyAtomCount	Number of Heavy atoms in monomer	natural number
NHOHCount	Number of NHs or OHs in monomer	natural number
NOCount	Number of Nitrogens and Oxygens in monomer	natural number
NumAliphaticCarbocycles	Number of aliphatic carbocycles in monomer	natural number
NumAliphaticHeterocycles	Number of aliphatic	natural number

	heterocycles in monomer	
NumAliphaticRings	Number of aliphatic rings	natural number
NumAromaticCarbocycles	Number of aromatic carbocycles in monomer	natural number
NumAromaticHeterocycles	Number of aromatic heterocycles in monomer	natural number
NumAromaticRings	Number of aromatic rings	natural number
NumHAcceptors	Number of Hydrogen Bond Acceptors in monomer	natural number
NumHDonors	Number of Hydrogen Bond Donors in monomer	natural number
NumHeteroatoms	Number of heteroatoms in monomer	natural number
NumRotatableBonds	Number of Rotatable Bonds in monomer	natural number
RingCount	Number of rings in monomer	natural number

To enable unconditional generation of polymer structures using the PolyTAO model, we adopted a two-step approach: (1) generating plausible descriptor values as inputs, and (2) obtaining polymer SMILES by feeding these values into the model.

Ensuring that the generated descriptor values accurately represent the real polymer population was critical. To achieve this, we derived empirical distributions for each descriptor from PolyInfo - the largest available database of experimentally validated polymers (Figure S1). For descriptors with discrete (natural number) values, we modeled their distributions using categorical distributions. The sole continuous descriptor, molecular weight (MolWt), was instead modeled using a log-normal distribution, given its strictly positive and right-skewed nature (Figure S2). The parameters of these distributions were estimated via maximum likelihood estimation (MLE), ensuring a statistically robust fit to the observed data.

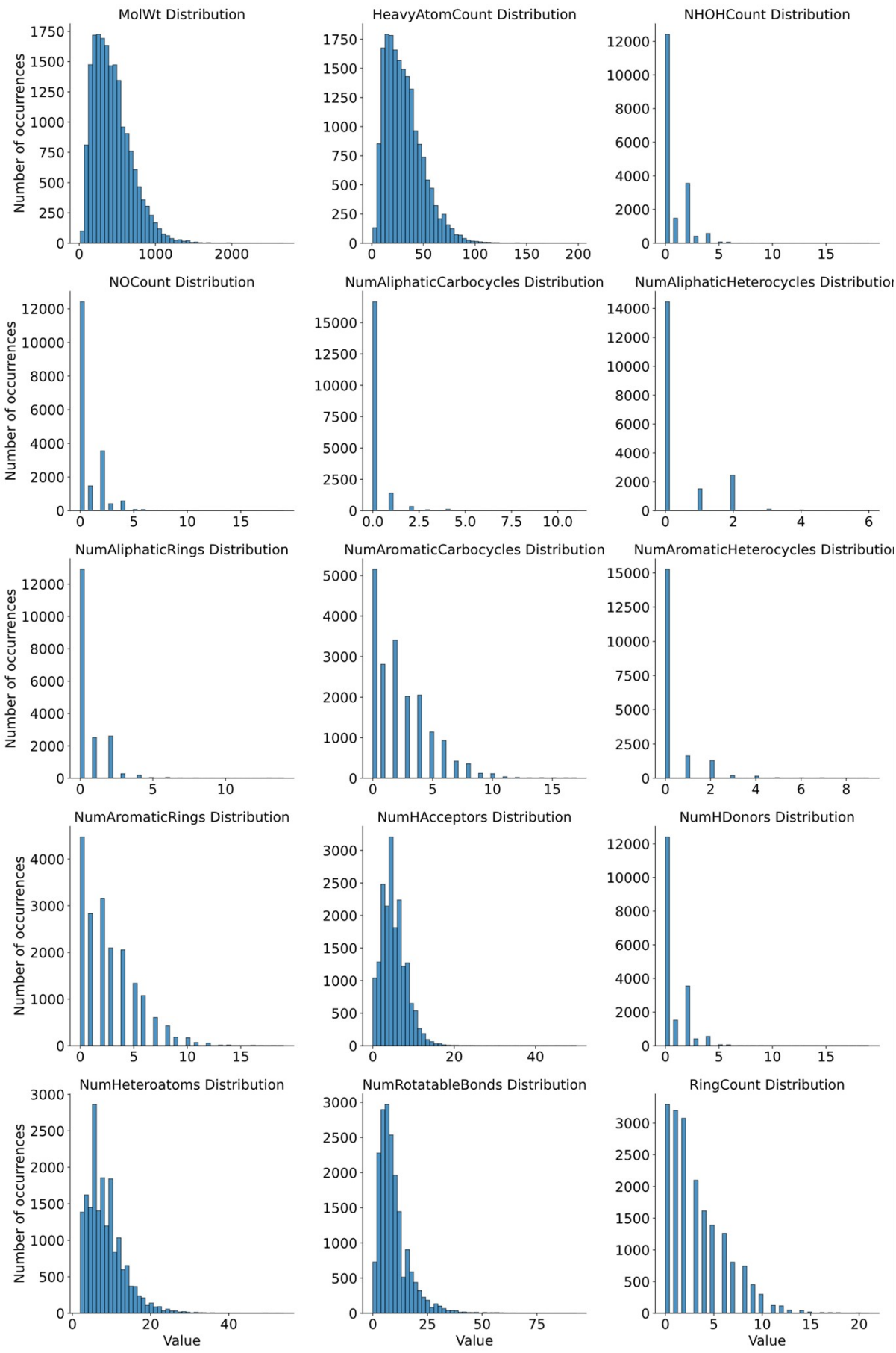

Figure S1. Distribution of PolyTAO descriptors in PolyInfo.

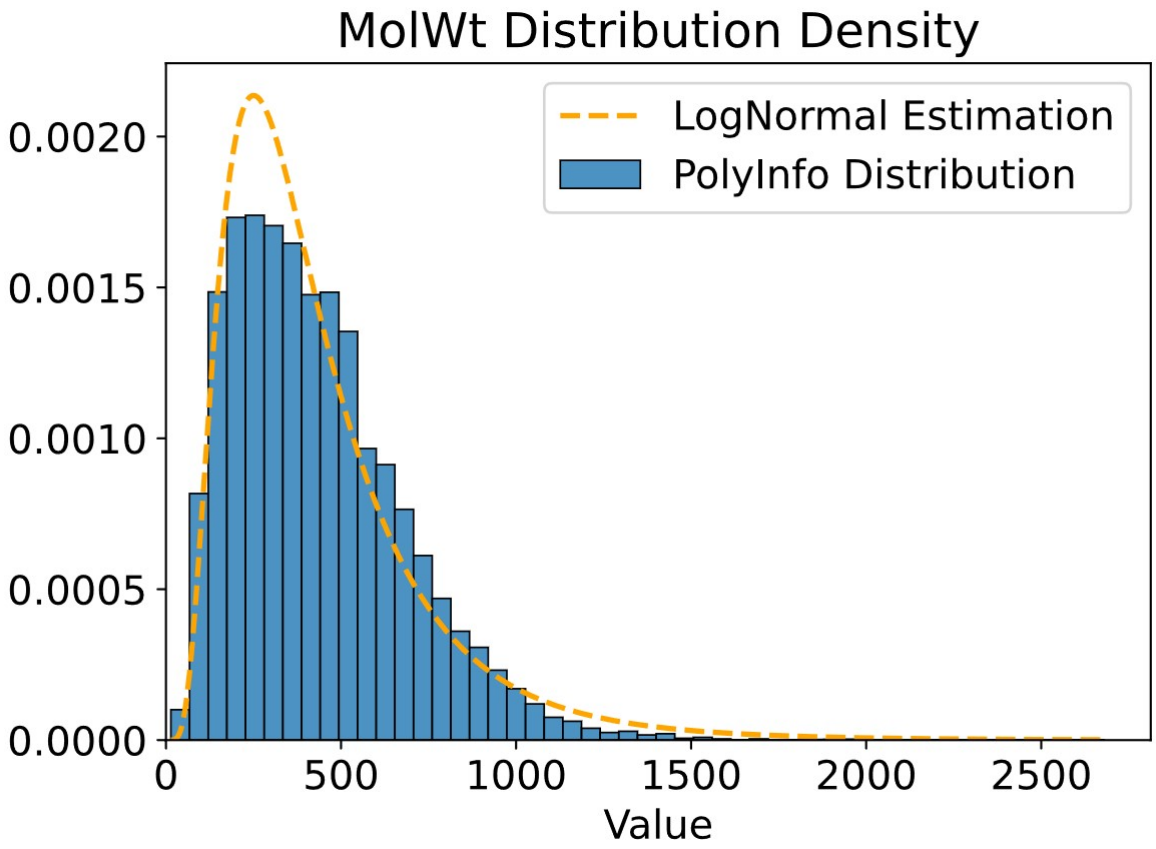

Figure S2. MolWt descriptor distribution from PolyInfo (blue bins) and its log-normal approximation (orange dashed line).

Following distribution parameter estimation, descriptor values were independently sampled from their marginal distributions to form inputs for the PolyTAO model. It is important to note, that we intentionally use independent sampling to avoid the risk of exposing explicit data leakage in the case of providing the PolyTAO model full parameters on joint distribution. This preserves the validity of using PolyTAO as pretraining data while respecting its design as a conditional, not unconditional, generator.

These inputs were then fed into the model to generate corresponding polymer SMILES strings. The sampling and generation process was iteratively repeated until a dataset of 1 million unique, valid polymer structures was obtained.

2.2 Data filtration

We define the improved validity metric for the polymer structure called “p-validity”. All SMILES strings used for model training satisfy described criteria.

3. SMILES Generators Fine-Tuning

Given the existence of pre-trained models for de-novo SMILES generation, which are trained on large datasets of bioactive molecules (e.g., ZINC, ChEMBL), a logical next step is to tune them for generating p-SMILES. Consequently, we selected SMILES-GPT,[4] a model based on the GPT-2 architecture pre-trained on the PubChem dataset, as our foundation model.

The primary objective of this generator is to propose novel polymer structures; therefore, its performance is critically dependent on the generation of unique and valid p-SMILES strings. Thus, the key evaluation metrics are the ratio of valid p-SMILES and the ratio of unique valid p-SMILES within the generated set.

Our initial approach involved a full fine-tuning of the SMILES-GPT model on a dataset of PolyInfo SMILES (which contains about 18k p-SMILES). This strategy yielded a notably low proportion of unique valid SMILES (0.2%), which was deemed insufficient for our purposes. Next, we employed Low-Rank Adaptation (LoRA)[5] to train adapters specifically for the c_attn and c_proj attention layers. This approach, however, resulted in further performance degradation.

We hypothesized that this failure could be attributed to the low frequency of the '*' symbol in the model's pre-training corpus. This symbol is essential in p-SMILES notation to denote the two endpoints of the polymer repeat unit but is rare in standard small-molecule SMILES. To address this potential bottleneck, we expanded the LoRA training to include the token embedding (wte) and language model head (lm_head) layers. Unfortunately, this modification did not lead to any improvement in performance.

A final experiment involved applying the same LoRA methodology using the larger PI1M corpus (about 1m p-SMILES). This approach proved significantly more successful, producing 37.6% unique p-valid p-SMILES. However, this model was outperformed by the PoGE model across all descriptor-based distribution metrics (see Extended Model Testing section).

A comprehensive summary of all fine-tuning experiments, including benchmark scores from the pre-trained PoGE model and its supervised fine-tuned (SFT) version, is presented in Table S2. It is worth noting that PoGE after pre-training stage produces significantly more unique polymers than PoGE after SFT, however, its distribution-based metrics are far from optimal Table S5.

Table S2. SMILES-GPT fine-tuning results

Model Description	Valid	p-valid	Unique valid	Unique p-valid
Full fine-tune on PolyInfo	0.150	0.002	0.150	0.002
LoRA for c proj and c attn layers on PolyInfo	0.068	0.008	0.068	0.008
LoRA for c proj, c attn, lm head and wte layers on PolyInfo	0.080	0.025	0.080	0.025
LoRA for c proj and c attn layers on PI1M	0.411	0.376	0.410	0.376
LoRA for c proj, c attn, lm head and wte layers on PI1M	0.020	0.018	0.020	0.018
Pretrain	0.833	0.830	**0.830**	**0.828**
SFT	**0.871**	**0.864**	0.652	0.647

4. Extended Model Testing

UMAP projections[6] of ECFP representations (Figure S3) reveal that PoGE populates regions adjacent to PolyInfo clusters, implying synthesizability via known pathways. In contrast, PolyTAO and PI1M exhibit fragmented distributions, reflecting their reliance on conditional generation and smaller training sets, respectively. These results highlight PoGE's ability to produce diverse, chemically coherent structures.

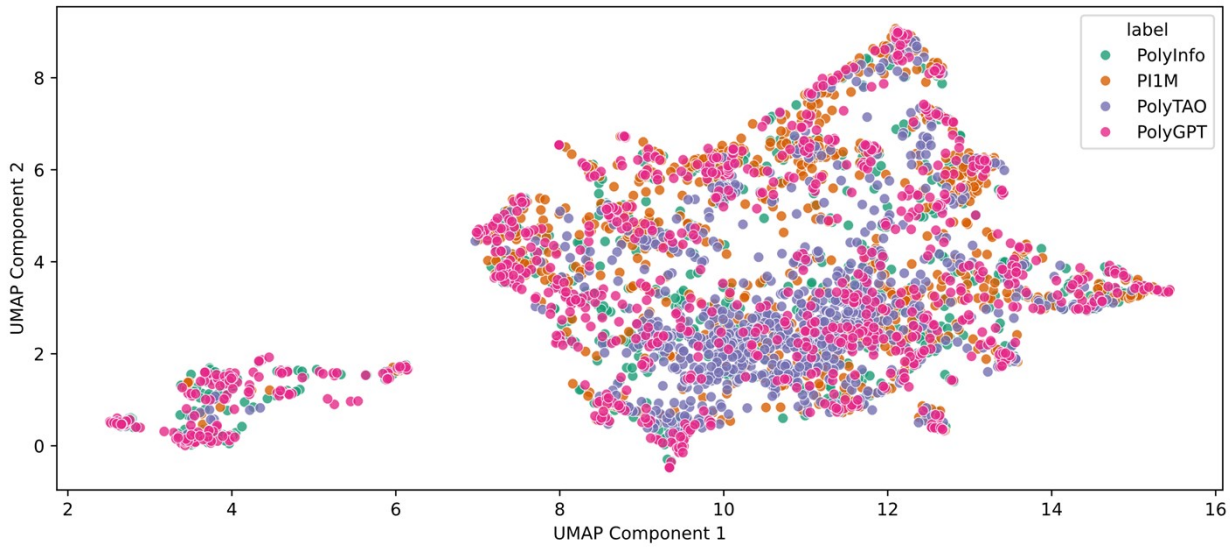

Figure S3. UMAP of ECFP representations of 1k samples of PolyInfo (green), PI1M (orange), PolyTAO (purple) and PoGE (pink)

Metrics based on BRICS fragments and Bemis–Murcko scaffolds are shown in Table S3.

Fragment similarity (BRICS) compares distributions of BRICS fragments in generated and reference sets. Denoting $c_f(A)$ a number of times a substructure f appears in molecules from set A, and a set of fragments that appear in either G or R as F, the metric is defined as a cosine similarity:

$$Frag(G,R) = \frac{\sum_{f \in F} [c_f(G) \cdot c_f(R)]}{\sqrt{\sum_{f \in F} c_f^2(G)} \cdot \sqrt{\sum_{f \in F} c_f^2(R)}} \#(1)$$

Scaffold similarity (Bemis–Murcko scaffolds) is similar to fragment similarity metric, but instead of fragments we compare frequencies of Bemis–Murcko scaffolds. We use MOSES implementation of this algorithm which additionally considers carbonyl groups attached to rings as part of a scaffold. Denoting $c_s(A)$ a number of times a scaffold s appears in molecules from set A, and a set of fragments that appear in either G or R as S, the metric is defined as a cosine similarity:

$$Frag(G,R) = \frac{\sum_{s \in S} [c_s(G) \cdot c_s(R)]}{\sqrt{\sum_{s \in S} c_s^2(G)} \cdot \sqrt{\sum_{s \in S} c_s^2(R)}} \#(2)$$

While a comparison of the BRICS and Bemis–Murcko scaffolds similarities is provided in Table S3, its utility is limited. For instance, PolyTAO achieves the best scores in BRICS and Bemis–Murcko scaffolds yet performs worst in the SNN metric against the PolyInfo dataset. This discrepancy suggests that the BRICS and Bemis–Murcko scaffold fragments, which are derived from common synthons in organic chemistry, may not be appropriate for evaluating polymer datasets, where the relevant chemical building blocks differ significantly.

Table S3. BRICS and scaffold similarities

Model Name	BRICS	Bemis–Murcko scaffolds
PI1M	0.009	0.201
PolyTAO	**0.107 ± 0.014**	**0.460 ± 0.038**

PoGE	0.004 ± 0.001	0.058 ± 0.011

As it was mentioned previously, distribuition-based-metrics were also computed via Jensen–Shannon divergence (Table S4). Moreover, here one could find the extended table with distribution-based-metrics computed via Wasserstein distance for the best models from smiles-gpt fine-tuning (Table S5).

Table S4. Extended descriptors-distribution-based metrics. (Jensen–Shannon divergence)

Model name	Molar mass	Aromatic fraction	Rotatable bond fraction	Heteroatom fraction	TPSA
PI1M	0.150±0.021	0.115±0.018	0.132±0.022	**0.043±0.007**	0.089±0.016
PolyTAO	0.359±0.061	0.268±0.047	0.104±0.018	0.172±0.031	0.148±0.025
PoGE (after SFT stage)	**0.060±0.007**	**0.076±0.009**	**0.062±0.008**	0.045±0.007	**0.038±0.005**
PoGE (after SFT stage)	0.158±0.028	0.272±0.046	0.189±0.032	0.050±0.009	0.089±0.016
SMILES-GPT LoRA for c_attn and c_proj on PI1M	0.180±0.032	0.351±0.063	0.288±0.046	0.123±0.021	0.124±0.021

Table S5. Extended descriptors-distribution-based metrics. (Wasserstein distance)

Model name	Molar mass	Aromatic fraction	Rotatable bond fraction	Heteroatom fraction	TPSA
PI1M	64	0.068	0.062	**0.010**	8.3
PolyTAO	195±24	0.108±0.012	0.038±0.005	0.048±0.004	17.5±0.8
PoGE (after SFT stage)	**33±2**	**0.026±0.002**	**0.030±0.002**	0.016±0.001	**4.6±0.3**
PoGE (after SFT stage)	77±8	0.128±0.020	0.096±0.006	0.011±0.002	8.9±0.5
SMILES-GPT LoRA for c_attn and c_proj on PI1M	120±15	0.160±0.022	0.118±0.016	0.029±0.002	19.3±1.2

In Figure S4, one can find examples of p-valid yet chemically implausible linear homopolymer structures.

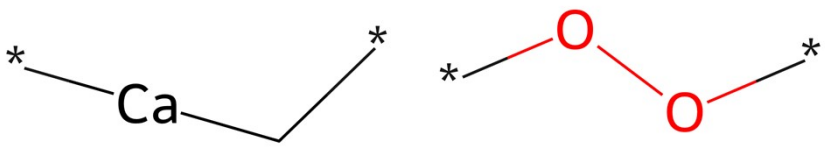

Figure S4. Example of chemically implausible p-valid structures.

In Figure S5 one could find examples of generated polymers.

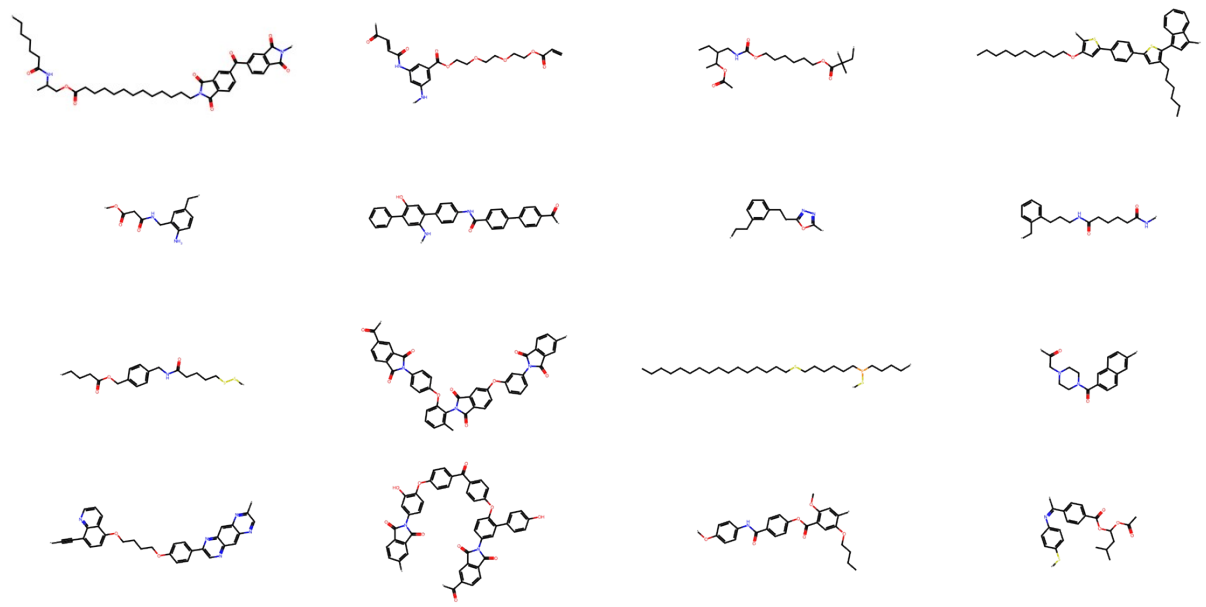

Figure S5. Examples of generated polymers

5. References

(1) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(2) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using "Drug-like" Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503–1507. https://doi.org/10.1002/cmdc.200800178.

(3) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39* (15), 2887–2893. https://doi.org/10.1021/jm9602928.

(4) Adilov, S. Generative Pre-Training from Molecules. ChemRxiv September 16, 2021. https://doi.org/10.26434/chemrxiv-2021-5fwjd.

(5) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; others. Lora: Low-Rank Adaptation of Large Language Models. *ICLR* **2022**, *1* (2), 3.

(6) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2020**, No. arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426.