

УДК 66.017

Физически - информированные большие языковые модели для de - novo генерации перспективных полимерных структур

Злобин Иван Сергеевич^(*)

zlobin@bmstu.ru

Беспалов Иван Андреевич

bespalovia@bmstu.ru,

Шнайдер Марк Юрьевич

schneidermyu@bmstu.ru

Аверочкин Глеб Михайлович

averochkin@bmstu.ru

Александров Евгений Викторович

alexandrov@bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. Разработка новых полимеров с заданными свойствами является ключевой задачей материаловедения. Генеративные модели, успешно применяемые для дизайна малых молекул, сталкиваются с существенными ограничениями при работе с полимерами из-за нехватки данных и неадекватности существующих методов валидации. В данной работе представлена трансформерная модель для генерации полимерных структур, основанная на архитектуре GPT-2. Мы вводим полимер-специфичные критерии валидации (p-validity) и предлагаем методологию оценки на основе расстояния Вассерштейна для сравнения распределений физико-химических дескрипторов. Модель превосходит существующие подходы PI1M и PolyTAO по ключевым метрикам, включая p-validity ($0,86 \pm 0,01$) и соответствие распределений молекулярной массы и других критических свойств реальным полимерным данным. Наша работа предоставляет воспроизводимый бенчмарк и открытые инструменты для дальнейших исследований в области обратного дизайна полимеров.

Ключевые слова: генеративные модели, полимеры, трансформер, SMILES, LLM, физически-информированное машинное обучение

Physics - Informed Large Language Models for De Novo Generation of Promising Polymer Structures

Zlobin Ivan Sergeevich^(*)

zlobin@bmstu.ru

Bespalov Ivan Andreevich

bespalovia@bmstu.ru

Shneider Mark Yurievich

schneidermyu@bmstu.ru

Averochkin Gleb Mikhailovich

averochkin@bmstu.ru

Alexandrov Evgeny Viktorovich

alexandrov@bmstu.ru

Bauman Moscow State Technical University, Moscow, Russia

Abstract. The development of new polymers with specified properties is a key challenge in materials science. Generative models, successfully used for small molecule design, face significant limitations when applied to polymers due to data scarcity and the inadequacy of existing validation methods. This work presents a transformer model for generating polymer structures, based on the GPT-2 architecture. We introduce polymer-specific validation

criteria (p-validity) and propose an evaluation methodology based on the Wasserstein distance for comparing distributions of physicochemical descriptors. The model outperforms existing approaches PI1M and PolyTAO in key metrics, including p-validity (0.86 ± 0.01) and the similarity of molecular weight distributions and other critical properties to real polymer data. Our work provides a reproducible benchmark and open-source tools for further research in the field of inverse polymer design.

Keywords: generative models, polymers, transformer, SMILES, LLM, physics-informed machine learning

Введение. Современные методы искусственного интеллекта, в частности генеративные модели, ускоряют процесс открытия новых материалов. Машинное обучение позволяет эффективно решать прямую задачу — предсказания свойств материалов по их структуре [1]. Трансформерные архитектуры, изначально разработанные для обработки естественного языка, показали высокую эффективность в генерации молекулярных структур, представленных в виде строк SMILES (Simplified Molecular Input Line Entry System) [2, 3]. Однако прямое применение этих подходов к дизайну полимеров, особенно в задачах обратного дизайна, когда требуется получить структуру с заданными макроскопическими свойствами (например, температурой стеклования или газопроницаемостью), сталкивается с уникальными вызовами.

Основное отличие заключается в данных. Для малых молекул существуют обширные базы данных, такие как PubChem [4] и ChEMBL [5], содержащие миллионы соединений. В области полимеров публичные базы данных, такие как PolyInfo [6], значительно меньше и содержат лишь десятки тысяч записей. Существующие генеративные модели для полимеров, такие как PI1M [7], основаны на рекуррентных нейронных сетях (RNN) и имеют ограничения в моделировании длинных зависимостей и сложных структурных мотивов. Другой подход, PolyTAO [8], использует условную генерацию на основе структурных дескрипторов, что не всегда интуитивно для исследователя, который оперирует целевыми макроскопическими свойствами.

Критической проблемой является и оценка качества сгенерированных полимерных структур. Стандартные бенчмарки (MOSES [9], guacaMOL [10]) и метрики (BRICS-фрагменты [11], скаффолды Бемиса-Мурко [12]), разработанные для лекарственных молекул, не учитывают специфику полимеров и могут давать вводящие в заблуждение результаты.

В данной работе мы представляем генеративную модель на основе трансформера, предназначенную для создания химически корректных и физико-химически правдоподобных полимерных структур. Наша работа включает:

- введение строгих полимер-специфичных критериев валидации (p-validity);
- разработку методологии оценки на основе сравнения распределений ключевых дескрипторов с использованием расстояния Вассерштейна.
- создание и обучение модели, превосходящей существующие аналоги по основным метрикам.

Методы и материалы. Подготовка данных и критерии валидности.

Для обучения модели был создан объединенный корпус, включающий:

PI1M [7]: ~1 млн синтетических полимерных структур в формате r-SMILES (специальный вариант SMILES для полимеров, где '*' обозначает концы повторяющегося звена).

корпус PolyTAO: ~1 млн структур, сгенерированных с помощью модели PolyTAO [8] на основе синтезированных дескрипторов, распределения которых аппроксимированы по данным PolyInfo.

PolyInfo [6]: 18 000 уникальных экспериментально подтвержденных структур, реплицированных 5 раз для увеличения веса высококачественных данных в процессе обучения.

Был введен строгий критерий **p-validity** для полимерных структур. Сгенерированная структура считается p-valid, если:

Строка является валидным SMILES (проходит конвертацию в RDKit MOL-объект).

Каждая точка полимеризации (*) соединена ровно с одним атомом.

Все связи в точках полимеризации имеют одинаковый тип.

В SMILES присутствуют ровно 2 точки полимеризации.

Архитектура модели и обучение. В качестве основы была выбрана архитектура GPT-2 [13]. Из-за ограниченного объема данных использовался уменьшенный вариант модели: 6 трансформерных слоев, размерность скрытого измерения — 256, 8 голов внимания. Все SMILES-строки были токенизированы с помощью Byte-Pair Encoding (BPE) [14] для эффективной работы с последовательностями.

Обучение проводилось в два этапа:

Предварительное обучение (Pre-training): Модель обучалась на объединенном корпусе (PI1M, PolyTAO Corpus, реплицированный PolyInfo) для изучения общих структурных паттернов.

Контролируемое тонкое дообучение (Supervised Fine-Tuning, SFT): Модель дообучалась исключительно на высококачественных данных из PolyInfo для улучшения генерации химически правдоподобных структур.

Параметры генерации (температура, top_p, top_k) были оптимизированы с помощью фреймворка Optuna [15] для максимизации доли уникальных p-valid полимеров.

Результаты. Для всесторонней оценки использовались два типа метрик:

Структурные метрики: Валидность (Valid), Полимерная валидность (p-valid), Уникальность (Unique), Новизна (Novelty), Сходство с ближайшим соседом (SNN), Внутреннее разнообразие (IntDiv) (табл. 1).

Метрики на основе дескрипторов: Для сравнения распределений ключевых физико-химических свойств (молярная масса, ароматичность, доля вращающихся связей, доля гетероатомов, топологическая полярная поверхность TPSA) между сгенерированными данными и PolyInfo использовалось расстояние Вассерштейна [16]. Эта метрика обеспечивает чувствительную оценку схожести распределений (табл. 2).

Таблица 1

Структурные метрики генерации

Модель	Valid	p-valid	Unique	Novelty	SNN	IntDiv
PI1M	0,80	N/A	N/A	N/A	0,63	0,86
PolyTAO	0,71±0,03	0,63±0,02	0,63±0,02	0,99±0,01	0,31±0,04	0,89±0,03
Наша модель	0,87±0,01	0,86±0,01	0,65±0,02	0,87±0,03	0,74±0,02	0,87±0,02

Таблица 2

Расстояние Вассерштейна для распределений дескрипторов

Модель	Молярная масса	Ароматичность	Доля вращ. связей	Доля гетероатомов	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
Наша модель	33±2	0,026±0,002	0,030±0,002	0,016±0,001	4,6±0,3

Обсуждение. Наша модель продемонстрировала превосходство над моделями PI1M и PolyTAO по всем основным структурным метрикам (табл. 1).

Модель генерирует 86 % полимер-валидных структур, что значительно выше, чем у аналогов. Высокий показатель SNN (0,74) указывает на близость сгенерированных структур к реальным данным из PolyInfo.

Анализ распределений дескрипторов подтвердил, что наша модель лучше всего воспроизводит статистические свойства реальных полимеров (табл. 2). По всем ключевым дескрипторам (молярная масса, ароматичность, доля вращающихся связей, TPSA) расстояние Вассерштейна для нашей модели оказалось значительно ниже, чем у конкурентов.

Попытки тонкой настройки (fine-tuning) предобученной модели для малых молекул SMILES-GPT [17] на полимерных данных (PolyInfo и PI1M) с использованием полной настройки и метода LoRA [18] не увенчались успехом. Наилучший результат (37,6 % уникальных p-valid полимеров) был достигнут при использовании LoRA на PI1M, однако эта модель значительно уступала лучшей модели по метрикам на основе дескрипторов.

Заключение. В данной работе представлен трансформерный фреймворк для генерации полимерных структур, обеспечивающий высокую химическую корректность и физико-химическое правдоподобие. Мы показали, что последовательное моделирование p-SMILES позволяет неявно учесть сложные структурные правила полимерной химии без явного задания дескрипторов.

Ключевыми достижениями являются:

- введение и формализация критерия p-validity для полимеров;
- разработка физически обоснованной схемы оценки на основе расстояния Вассерштейна для распределений дескрипторов;

– создание модели, превосходящей существующие аналоги, и обеспечение полной воспроизводимости результатов за счет публикации кода, данных и бенчмарка.

Наша работа создает основу для решения задач условной генерации и обратного дизайна полимеров с целевыми свойствами для таких приложений, как гибкая электроника и мембраны для разделения газов. Дальнейшая работа будет сосредоточена на разработке полимер-специфичных метрик синтетической доступности.

Работа выполнена в рамках программы государственной поддержки центров Национальной технологической инициативы на базе образовательных организаций высшего образования и научных организаций (Центр НТИ «Цифровое материаловедение: новые материалы и вещества» при МГТУ им. Н.Э. Баумана).

СПИСОК ИСТОЧНИКОВ

- [1] Zlobin I., Toroptsev N., Averochkin G. et al. Pre-trained Mol2Vec Embeddings as a Tool for Predicting Polymer Properties. *Chin J Polym Sci.*, 2014, vol. 42, pp. 2059–2068. <https://doi.org/10.1007/s10118-024-3237-y>
- [2] Arús-Pous J. et al. SMILES-based deep generative scaffold decorator for de-novo drug design. *Journal of cheminformatics*, 2020, vol. 12, no. 1, art. no. 38. <https://doi.org/10.1186/s13321-020-00441-8>
- [3] Bagal V. et al. MolGPT: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*, 2021, vol. 62, no. 9, pp. 2064–2076. <https://doi.org/10.1021/acs.jcim.1c00600>
- [4] Kim S. et al. PubChem 2025 update. *Nucleic acids research*, 2025, vol. 53, no. D1, pp. D1516–D1525. <https://doi.org/10.1093/nar/gkae1059>
- [5] Mendez D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 2019, vol. 47, no. D1, pp. D930–D940. <https://doi.org/10.1093/nar/gky1075>
- [6] Otsuka S. et al. PoLyInfo: Polymer database for polymeric materials design. *International conference on emerging intelligent data and web technologies, IEEE*, 2011, pp. 22–29. <https://doi.org/10.1109/EIDWT.2011.13>
- [7] Ma R., Luo T. PI1M: a benchmark database for polymer informatics. *Journal of Chemical Information and Modeling*, 2020, vol. 60, no. 10, pp. 4684–4690. <https://doi.org/10.1021/acs.jcim.0c00726>
- [8] Qiu H., Sun Z.Y. On-demand reverse design of polymers with PolyTAO. *npj Computational Materials*, 2024, vol. 10, no. 1, art. no. 273. <https://doi.org/10.1038/s41524-024-01466-5>
- [9] Polykovskiy D. et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 2020, vol. 11, art. no. 565644. <https://doi.org/10.3389/fphar.2020.565644>
- [10] Brown N. et al. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 2019, vol. 59, no. 3, pp. 1096–1108. <https://doi.org/10.1021/acs.jcim.8b00839>
- [11] Degen J. et al. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 2008, vol. 3, no. 10, art. no. 1503. <https://doi.org/10.1002/cmdc.200800178>
- [12] Bemis G.W., Murcko M.A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry*, 1996, vol. 39, no. 15, pp. 2887–2893. <https://doi.org/10.1021/jm9602928>

-
- [13] Radford A. et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, vol. 1, no. 8, art. no. 9.
 - [14] Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units. Arxiv preprint arXiv:1508.07909, 2015. <https://doi.org/10.48550/arXiv.1508.07909>
 - [15] Akiba T. et al. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631. <https://doi.org/10.1145/3292500.3330701>
 - [16] Panaretos V.M., Zemel Y. Statistical aspects of Wasserstein distances. Annual review of statistics and its application, 2019, vol. 6, no. 1, pp. 405–431. <https://doi.org/10.1146/annurev-statistics-030718-104938>
 - [17] Adilov S. Generative pre-training from molecules. 2021. <https://doi.org/10.26434/chemrxiv-2021-5fwjd>
 - [18] Hu E.J. et al. Lora: Low-rank adaptation of large language models. ICLR, 2022, vol. 1, no. 2, art. no. 3. <https://doi.org/10.48550/arXiv.2106.09685>