

Машинное обучение с физически обоснованными интерпретируемыми дескрипторами для предсказания свойств полимеров и дизайна мономеров

Аверочкин Глеб Михайлович

averochkin@bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. Надежное предсказание свойств полимеров по структуре мономера требует дескрипторов, которые обеспечивают точность и интерпретируемость прогнозов. В данной работе мы представляем физически обоснованную QSPR-концепцию, объединяющую атомные Ван-дер-Ваальсовы объемы, подсчет фрагментов, статистику связей, квантовохимические характеристики и отобранные дескрипторы RDKit. Используя мономеры из базы данных PolyInfo и соответствующие свойства полимеров: плотность, температуру стеклования, плавления и термического разложения, была достигнута высокая точность моделей регрессии гауссовых процессов (R^2 до 0,89). Анализ причинных связей выявил дескрипторы Ван-дер-Ваальсовых объемов как определяющие структурные факторы, в то время как анализ SHAP показал согласованную роль ароматичности, полярности и гибкости молекулы. По сравнению с традиционными признаками RDKit, наши дескрипторы обеспечивают существенно более высокую интерпретируемость и дают практические ориентиры для рационального проектирования мономеров. Для воспроизводимой и интерпретируемой работы в области полимерной информатики создан открытый веб-инструмент для фичеризации мономеров и предсказания свойств полимеров.

Ключевые слова: полимерная информатика, анализ связей структура-свойство, машинное обучение

Machine Learning with Physically Grounded, Interpretable Descriptors for Polymer Property Prediction and Monomer Design

Averochkin Gleb Mihailovich

averochkin@bmstu.ru

BMSTU, Moscow, Russia

Abstract. Reliable prediction of polymer properties from monomer structure requires descriptors that are both accurate and interpretable. Here we introduce a physically grounded QSPR framework combining atomic van der Waals volumes, fragment counts, bonding statistics, quantum-chemical features, and selected RDKit descriptors. Using monomers from PolyInfo database with reported density, glass transition, melting, and thermal decomposition temperatures, Gaussian Process Regression models achieved high accuracy (R^2 up to 0.89) with calibrated uncertainty. Causal inference analysis identified Van der Waals volume descriptors as the dominant structural drivers, while SHAP analysis revealed consistent roles of aromatic content, polarity, and molecular flexibility. Compared with conventional RDKit features, our descriptors offer substantially greater interpretability.

ity and actionable guidance for rational monomer design. A public web tool is provided for featurization and property prediction, enabling reproducible and interpretable polymer informatics workflows.

Keywords: polymer informatics, quantitative structure–property relationships, machine learning

Введение. Точное и интерпретируемое предсказание свойств полимеров, исходя из молекулярной структуры мономеров, является одним из ключевых вызовов для полимерной информатики сегодня [1–4]. Несмотря на недавние успехи машинного обучения, наиболее часто используемым представлениям молекул — такие как отпечатки Моргана [5, 6], эмбединги mol2vec [7, 8], топологические дескрипторы [3, 9], и эмбединги графовых нейронных сетей [10, 11] — недостает физической обоснованности и химической интерпретируемости [12]. Поэтому, они плохо подходят для рационального дизайна мономеров для достижения желаемых свойств (обратной задачи), в котором необходимо понимание влияния структурных модификаций на целевое свойство материала [13].

В то же время задача предсказания свойств полимеров по структуре мономеров была поставлена задолго до проникновения машинного обучения в полимерную информатику. Аскадский [14, 15] разработал метод инкрементов — аддитивную модель для свойств полимеров, выражающую их через Ван-дер-Ваальсовы объемы атомов, составляющих мономер. Она была получена с помощью нескольких физических допущений в решении уравнения Шредингера в приближении гармонического осциллятора. Эта аддитивная схема была использована для предсказания широкого круга свойств полимеров с использованием небольших размеров датасетов, в основном внутри одного класса полимеров [16–18]. Впоследствии, Аскадский расширил эту методологию и создал программу КАСКАД [14, 15], работающую по принципу аддитивных схем. Эта программа была недавно применена для создания датасета для машинного обучения [10]. При этом, у этого подхода остаются серьезные ограничения: невозможность улавливать нелинейные зависимости, небольшое количество доступных дескрипторов, и отсутствие открытого доступа к программе КАСКАД. Другой подход к анализу связей структуры и свойств в полимерах был разработан Бицерано. Этот подход также основан на аддитивных схемах молекулярных подструктур, но вместо инкрементов атомов использует вклады структурных фрагментов и функциональных групп. Эта методология, описанная Бицерано в книге «Prediction of Polymer Properties» [19], была применена для предсказания термических и механических свойств полимеров. В отличие от подхода Аскадского, основанного на физических допущениях, модель Бицерано является полностью импирической и полагается на статистические корреляции между дескрипторами структуры и физическими свойствами. При этом значительная часть этих дескрипторов является топологическими, а значит, слабоинтерпретируемыми. Недавние разработки Бицерано были включены в платформу Polymer Genome, к которой тоже нет открытого доступа.

Таким образом, исследователь, стремящийся проанализировать датасет полимерных структур с использованием современных подходов data science и полимерной информатики, сталкивается с серьезным ограничением: среди доступных способов фичеризации структур мономеров в физически осмысленные численные дескрипторы, ни один не спроектирован специально для полимеров. Широко используемые хемоинформатические библиотеки, такие как RDKit и Mordred были разработаны для малых молекул, и в них недостает дескрипторов для полимеров. Графовые нейросети недавно показали высокую точность предсказаний свойств полимеров. Тем не менее, эти техники имеют много ограничений в интерпретируемости и вычислительной сложности, когда применяются к задачам полимерной информатики. В то же время, программное обеспечение для работы со структурами мономеров КАСКАД и Polymer Genome не имеют открытого доступа, препятствуя их использованию в открытых и воспроизводимых исследованиях.

Материалы и методы. В качестве исходных данных использовались записи из базы PolyInfo, содержащие структуры мономеров и экспериментальные значения четырех ключевых свойств полимеров: плотности, температуры стеклования, температуры плавления и температуры начала термического разложения. Размеры подмножеств составили от 1500 до 7700 наблюдений в зависимости от свойства. Для каждого мономера структура была представлена в формате SMILES с обозначением сайтов полимеризации.

Формирование признаков опиралось на физически обоснованную схему, сочетающую несколько источников информации. Были рассчитаны объемные параметры атомов и групп по методологии Аскадского, позволяющей учитывать вклады различных типов межмолекулярных взаимодействий. Дополнительно учитывались структурные фрагменты и функциональные группы в духе подхода Бицерано, а также статистика типов химических связей, отражающая степень жесткости или гибкости основной и боковых цепей. Для повышения полноты описания были добавлены интерпретируемые молекулярные дескрипторы из библиотеки RDKit, такие как топологическая полярная поверхность, число вращаемых связей. Электронные и энергетические характеристики извлекались с помощью полуэмпирических квантово-химических расчетов методом GFN2-xTB, включая полную энергию, уровни HOMO и LUMO, величину дипольного момента и распределение зарядов Малликена.

Результаты. Разработанная система дескрипторов позволила построить предсказательные модели для четырех целевых свойств полимеров — плотности, температуры стеклования, температуры плавления и температуры термического разложения — на основе структур мономеров из базы PolyInfo. Применение моделей гауссовских процессов обеспечило высокую точность: коэффициент детерминации достигал 0,89 при предсказании температуры стеклования, 0,83 для плотности, около 0,71 для температуры плавления и более 0,71 для температуры начала термического разложения. Средняя абсолютная ошибка при этом составляла 25–45 °C для температурных свойств и около 0,05 г/см³ для плотности. Эти показатели сопоставимы или превосхо-

дят результаты, полученные при использовании стандартных дескрипторов RDKit или отпечатков Моргана.

Анализ причинных связей с использованием метода Double Machine Learning показал, что ведущую роль во всех четырех задачах играют дескрипторы Ван-дер-Ваальсовых объемов атомных окружений, сгруппированных по типу цепи и типу межмолекулярных взаимодействий. Например, увеличение объема боковой цепи связано с уменьшением плотности и ростом температуры стеклования, а вклад ароматических атомов в основной цепи положительно коррелирует с повышением T_g , температуры плавления и термического разложения. Дополнительное влияние оказывают параметры, связанные с полярностью и дипольными взаимодействиями, а также меры гибкости, такие как число вращаемых связей и фракция sp^3 -гибридизованных атомов. SHAP-анализ подтвердил эти закономерности, показав согласованность вклада одних и тех же признаков в предсказания разных свойств.

Обсуждение. Полученные результаты подтверждают, что физически обоснованная схема фичеризации позволяет воспроизводить ключевые закономерности структуры и свойств полимеров. При этом новые дескрипторы обеспечивают более широкий охват химических элементов и более подробное описание структуры главной и боковой цепи, совместимость с современными алгоритмами машинного обучения.

Сравнение с моделями на основе RDKit-дескрипторов и молекулярных отпечатков показало, что выигрыш в точности не всегда драматичен, однако ключевое преимущество новой методологии заключается в интерпретируемости. В то время как топологические индексы RDKit, такие как VSA или BCUT, трудно напрямую связать с конкретными структурными модификациями, наши дескрипторы позволяют ясно объяснить, какие именно элементы молекулы влияют на свойство и в каком направлении. Это открывает возможность не только предсказывать значения, но и использовать модель в обратной задаче рационального дизайна мономеров. Ограничения подхода связаны с тем, что в анализ включены только структурные признаки мономера, тогда как такие факторы, как тактичность, степень кристалличности или распределение молекулярных масс, также могут оказывать сильное влияние на свойства. Тем не менее, уже на уровне мономеров достигнуты показатели, достаточные для предварительного скрининга, что делает методологию практически полезной в задачах ускоренного поиска новых материалов.

Заключение. В работе проведен систематический анализ основных подходов к QSPR-моделированию полимеров с акцентом на пять семейств дескрипторов, показавший преимущество объемных характеристик по Аскадскому в точности прогнозов и причинных связях. На основе этих результатов создан специализированный пакет для фичеризации мономеров с веб-доступом и предобученными моделями для четырех свойств: плотности, температур стеклования, плавления и разложения. Применение причинного анализа и SHAP подтвердило ключевую роль Ван-дер-Ваальсовых объемов, а также влияние ароматичности, полярности и гибкости цепей, при этом жесткие и по-

лярные фрагменты повышают стабильность, а гибкие звенья ее снижают. Среди алгоритмов наилучший баланс точности и интерпретируемости обеспечили модели гауссовских процессов, давшие достоверные предсказания с корректной оценкой неопределенности. Сравнение с дескрипторами RDKit показало, что предложенный набор обеспечивает не только сопоставимую или более высокую точность, но и существенно большую интерпретируемость, позволяя напрямую связывать свойства с конкретными структурными элементами и использовать модель в задачах рационального дизайна мономеров.

Работа выполнена при поддержке Российского научного фонда, проект № 25-73-00026.

Дополнительная благодарность выражается за предоставление вычислительных ресурсов, необходимых для реализации проекта. Вычисления проведены в рамках программы государственной поддержки центров Национальной технологической инициативы (НТИ) на базе образовательных и научных организаций (Центр НТИ «Цифровое материаловедение: новые материалы и вещества» на базе МГТУ им. Н.Э. Баумана).

СПИСОК ИСТОЧНИКОВ

- [1] Ramprasad R., Batra R., Pilia G., Mannodi-Kanakkithodi A., Kim C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *Npj Comput. Mater.*, 2017, vol. 3, no. 54. <https://doi.org/10.1038/s41524-017-0056-5>
- [2] Audus D.J., De Pablo J.J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.*, 2017, vol. 6, pp. 1078–1082. <https://doi.org/10.1021/acsmacrolett.7b00228>
- [3] Zhao Y., Mulder R.J., Houshyar S., Le T.C. A review on the application of molecular descriptors and machine learning in polymer design. *Polym. Chem.*, 2023, vol. 14, pp. 3325–3346. <https://doi.org/10.1039/D3PY00395G>
- [4] Martin T.B., Audus D.J. Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polym. Au*, 2023, vol. 3, pp. 239–258. <https://doi.org/10.1021/acspolymersau.2c00053>
- [5] Zhong S., Guan X. Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties. *Environ. Sci. Technol.*, 2023, vol. 57, pp. 18193–18202. <https://doi.org/10.1021/acs.est.3c02198>
- [6] Rogers D., Hahn M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 2010, vol. 50, pp. 742–754. <https://doi.org/10.1021/ci100050t>
- [7] Jaeger S., Fulle S., Turk S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.*, 2018, vol. 58, pp. 27–35. <https://doi.org/10.1021/acs.jcim.7b00616>
- [8] Zlobin I., Toroptsev N., Averochkin G., Pavlov A. Pre-trained Mol2Vec Embeddings as a Tool for Predicting Polymer Properties. *Chin. J. Polym. Sci.*, 2024, vol. 42, pp. 2059–2068. <https://doi.org/10.1007/s10118-024-3237-y>
- [9] Balaban A.T. Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR, Similarity/Dissimilarity and Drug Design. *SAR QSAR Environ. Res.*, 1998, vol. 8, pp. 1–21. <https://doi.org/10.1080/10629369808033259>
- [10] Volgin I.V., Batyr P.A., Matseevich A.V., Dobrovskiy A.Yu., Andreeva M.V., Nazarychev V.M., Larin S.V., Goikhman M.Ya., Vizilter Y.V., Askadskii A.A., Lyulin S.V. Machine Learning with Enormous “Synthetic” Data Sets: Predicting Glass Transition Temperature of Polyimides Using Graph Convolutional Neural Networks. *ACS Omega*, 2022, vol. 7, pp. 43678–43691. <https://doi.org/10.1021/acsomega.2c04649>

- [11] Gurnani R., Kuenneth C., Toland A., Ramprasad R. Polymer Informatics at Scale with Multitask Graph Neural Networks. *Chem. Mater.*, 2023, vol. 35, pp. 1560–1567. <https://doi.org/10.1021/acs.chemmater.2c02991>
- [12] Gallegos L.C., Luchini G., St. John P.C., Kim S., Paton R.S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.*, 2021, vol. 54, pp. 827–836. <https://doi.org/10.1021/acs.accounts.0c00745>
- [13] Mikulskis P., Alexander M.R., Winkler D.A. Toward Interpretable Machine Learning Models for Materials Discovery. *Adv. Intell. Syst.*, 2019, vol. 1, art. no. 1900045. <https://doi.org/10.1002/aisy.201900045>
- [14] Askadskii A.A. *Computational Materials Science of Polymers*. Cambridge, Cambridge International Science Publishing, 2003.
- [15] Askadskii A. Methods for calculating the physical properties of polymers. *Rev. J. Chem.*, 2015, vol. 5, pp. 83–142. <https://doi.org/10.1134/S2079978015020016>
- [16] Askadskii A.A., Slonimskii G.L., Matveyev Yu.I., Korshak V.V. Molecular packing and the glass transition temperature of polymers. *Polym. Sci. USSR*, 1976, vol. 18, pp. 2363–2372. [https://doi.org/10.1016/0032-3950\(76\)90114-3](https://doi.org/10.1016/0032-3950(76)90114-3)
- [17] Askadskii A.A., Slonimskii G.L. Structure and Properties of Heat-resistant Polymers. *Russ. Chem. Rev.*, 1975, vol. 44, art. no. 767. <https://doi.org/10.1070/RC1975v044n09ABEH002375>
- [18] Askadskii A.A. Influence of Chemical Structure on the Properties of Polymers. *Pure and Applied Chemistry*, 1976, vol. 46, pp. 19–27. <https://doi.org/10.1351/pac197646010019>
- [19] Bicerano J. *Prediction of Polymer Properties*. CRC Press, Boca Raton, 2002.